TURING 图灵数学·统计学丛书 30

WILEY



北美精算师
考试指定参考书

Loss Models
From Data to Decisions

损失模型
从数据到决策

（第 2 版）

[美] Stuart A. Klugman
[加] Harry H. Panjer 著
[加] Gordon E. Willmot
吴岚 译



人民邮电出版社
POSTS & TELECOM PRESS

损失模型 从数据到决策

Loss Models From Data to Decisions

“不愧为经典之作。” ——《国际统计协会学报》

“这是一部杰作，每章都包含大量的实例和习题，几乎涵盖了统计领域内的所有相关主题。”

—— *Mathematical Reviews*

本书是精算领域的一部经典著作，也是北美精算师协会（SOA）和北美产险精算师协会（CAS）考试的指定参考书，被国内外众多著名高校采用为教材或者教学参考书。

书中全面探讨了精算损失模型和精算建模方法，并创造性地将概率模型和统计建模有机地结合起来，其中大量的实证案例分析，更有助于读者理解如何将精算理论运用于保险实务。

Stuart A. Klugman 著名统计学家。美国德雷克（Drake）大学精算学教授，SOA（北美精算师协会）会士，并曾任该协会副主席（2001–2003）。

Harry H. Panjer 著名统计学家。加拿大滑铁卢大学统计精算系荣休教授，曾任加拿大精算学会主席（1997–1998），SOA主席（2002–2003）。

Gordon E. Willmot 加拿大滑铁卢大学统计精算系教授，是国际风险理论研究方面的著名学者，曾担任SOA精算考试相关课程的建设工作。



WILEY

www.wiley.com

本书相关信息请访问：图灵网站 http://www.turingbook.com

读者热线：(010)88593802

反馈/投稿/推荐信箱：contact@turingbook.com

分类建议 数学/应用统计

人民邮电出版社网址 www.ptpress.com.cn





ISBN 978-7-115-19043-7



ISBN 978-7-115-19043-7/O1

定价：89.00 元

TURING 图灵数学·统计学丛书 30



北美精算师
考试指定参考书

Loss Models
Fr[illegible]ecisions

损失模型
从数据到决策
（第 2 版）

[美] Stuart A. Klugman
[加] Harry H. Panjer 著
[加] Gordon E. Willmot
吴岚 译

人民邮电出版社
北 京

图书在版编目（CIP）数据

损失模型：从数据到决策：第2版/(美)克卢格曼(Klugman, S. A.), (加)潘耶(Panjer, H. H.), (加)威尔莫特(Willmot, G. E.)著；吴岚译. —北京：人民邮电出版社, 2009. 1
（图灵数学·统计学丛书）
书名原文：Loss Models：From Data to Decisions
ISBN 978-7-115-19043-7/O1
Ⅰ.损… Ⅱ.①克…②潘…③威…④吴… Ⅲ.精算学
Ⅳ. F224.0
中国版本图书馆 CIP 数据核字（2008）第 165558 号

内 容 提 要

本书全面讨论了精算损失模型和精算建模方法, 共分5个部分. 第2部分至第5部分是全书的核心, 汇总了精算模型和精算建模方法2个体系的内容. 第2部分除介绍一般损失模型常用的概率分布外, 还介绍了保险精算中最基本的索赔频率模型、索赔额模型以及总损失模型, 并在此基础上讨论了破产理论模型. 随后3个部分的核心主题是精算建模方法, 从经验建模方法到参数化(统计)建模, 直至最后第5部分的模型修正方法和随机模拟方法.

本书是北美精算考试当前考试体系课程MLC和C的指定参考书, 是从事金融和精算工作的专业人士很有价值的参考书, 也可作为高等学校金融和精算方向相关课程的参考教材.

图灵数学·统计学丛书

损失模型:从数据到决策(第2版)

◆ 著　[美] Stuart A. Klugman　[加] Harry H. Panjer
　　[加] Gordon E. Willmot
　译　吴　岚
　责任编辑　明永玲
　执行编辑　边晓娜
◆ 人民邮电出版社出版发行　北京市崇文区夕照寺街14号
　邮编 100061　电子函件 315@ptpress.com.cn
　网址：http://www.ptpress.com.cn
　北京隆昌伟业印刷有限公司印刷
◆ 开本：700×1000　1/16
　印张：36
　字数：744千字　2009年1月第1版
　印数：1–3 000册　2009年1月北京第1次印刷

著作权合同登记号　图字：01-2007-5302号

ISBN 978-7-115-19043-7/O1

定价：89.00元

读者服务热线：(010)88593802　印装质量热线：(010) 67129223

反盗版热线：(010) 67171154

译者简介

吴岚 北京大学数学科学学院金融数学系副教授, 中国精算学会会员, 北京大学博士 (数理统计专业) 毕业.

1990 年至今在北京大学数学科学学院任教, 主要讲授课程:《风险理论》、《金融统计方法》. 1997 年开始从事金融数学与精算学的教学和科研工作, 参加国家自然科学基金、国家科技部 973 项目等相关的研究工作, 并参与保险行业偿付能力监管标准方面的技术工作以及中国精算协会的精算教育方面的工作.

主要研究方向为金融风险管理与精算学. 具体的研究领域: 投资连结的寿险产品的定价和风险管理、保险公司资产负债管理技术、商业银行信用风险模型、金融机构监管的风险资本模型等.

译 者 序

精算学是植根于保险实践的一门学科. 它以数学、统计学的相关理论和方法为基础, 以解决保险实践中的定量化问题为研究目的. 20 世纪 80 年代之前, 精算学在西方主要涉及保险产品定价和准备金评估技术两个方面. 这方面工作的重要理论基础相对比较简单, 概括地说就是概率论中的大数定律. 有了这个基础, 只要达到了一定的承保规模, 很多风险都是可以忽略的. 也就是说, 很多问题都可以在确定性空间下解决. 与此同时, 欧洲大陆一些从事概率论理论研究的学者则从随机过程应用的角度研究保险经营的整体风险模型问题, 集中的表现是所谓的破产理论 (ruin theory). 这方面的理论研究更像是科学家在实验室里模拟现实社会, 是一种经过对实际保险业务经营进行简化、抽象后的分析, 所给出的解决方案至多是对现实决策的一种辅助的支持, 大多时候并不被业界采用.

进入 20 世纪 80 年代, 随着整体经济和金融环境的变化, 保险业也发生了巨大的变化, 最为突出的是除了承保风险外, 保险经营中其他风险的特质越来越复杂, 同时越来越需要一些基于整体层面的定量分析理论和技术方法.

自改革开放以来, 我国的保险业迅速发展, 作为保险业核心的精算技术也经历了从无到有的过程. 而且我国的宏观经济环境和金融环境的快速变革, 带来了保险产品和经营模式的瞬间多样化和持续不断的变化, 这些背景对中国精算职业和精算教育的影响都是巨大的. 对保险业务中的各种损失建立科学量化的模型是精算工作的基础, 随着现代概率统计学科的发展, 这方面的理论和方法也有很大的进展, 特别是统计建模技术更是随着计算机和各种计算方法的更新而不断进步.

本书也适应了上述潮流, 只用 1/3 的篇幅 (第 2 章到第 8 章) 介绍常见的精算模型, 除一般损失模型常用的概率分布外, 还有保险精算中最基本的索赔频率和索赔额模型以及总损失模型, 并在此基础上讨论了破产理论模型. 然后用近 2/3 的篇幅介绍精算建模方法. 这里涉及基本的数理统计方法、各种参数模型的估计、模型选择技术, 并且进一步给出了统计建模中更高层面的方法和技术, 也就是如何对已有的估计和模型进行调整和修正的问题. 这包括传统的插值、平滑等模型修匀方法和具有精算自身特点的信度理论和方法. 最后一章介绍随机模拟方法及其在精算建模中的应用.

本书的三位作者都享誉国际精算界. Harry H. Panjer 是国际著名的精算教授, 曾任北美精算师协会 (Society of Actuaries) 主席 (2002—2003). 他不仅在风险理论的理论研究上很有造诣, 而且在精算教育和精算职业发展方面也做了很多有价值的工作. 特别值得一提的是, Panjer 教授对中国的精算教育也非常热心, 他曾是北美精算协会为支持我国精算教育而参与南开精算硕士班教学的教授之一. 我有幸多次与 Panjer 教授进行过交流, 感觉他是一位具有很深的理论研究功底和很好的实务

感觉的学术领军式人物. Stuart A. Klugman 教授是经典教材《损失分布》([59]) 的主要作者. 他在保险损失建模方面做了大量的研究工作, 积累了丰富的经验, 他对本书的主要贡献是建模部分. Gordon E. Willmot 是加拿大滑铁卢大学 (University of Waterloo) 统计精算系的教授, 他在风险理论模型方面有很深的造诣, 也是经典教材《保险风险模型》([106]) 的作者之一. 他长期从事风险模型方面的理论研究, 在复合分布的算法、破产模型的计算等方面做出很好的成绩. 他在本书中的主要工作是精算模型部分, 特别是总损失模型和破产模型, 同时他还对信度理论有一定的研究, 北美精算考试关于信度理论的材料就出自 Willmot 教授之手.

本书在精算教材历史上第一次将精算技术中的数学模型 (以概率论、数值计算) 与数据建模 (统计建模) 有机地结合起来. 依译者之拙见, 本书有以下几个方面的特点: (1) 将保险精算概率模型和统计建模的所有内容有机地综合在一起; (2) 大量的实证案例分析, 非常有助于学生理解如何将精算理论与保险实务中的实践结合; (3) 文笔简洁, 定义和概念的叙述非常清楚明确.

本书适用于以下几个方面: (1) 作为高等院校精算专业课程"损失模型"和"风险理论"的主要参考书; (2) 作为参加北美精算考试课程 MLC 和 C 的中文参考书, 这将有助于考生理解英文原著, 提高复习的效率; (3) 作为参加中国精算师资格考试课程 05 和 08 的参考书; (4) 为业界精算实践的损失经验分析工作提供一个辅助工具, 特别是本书的第五部分在这方面非常有价值.

无论是从篇幅还是从内容来说, 本书都是一本经典著作. 因此, 为了尊重原书的价值和品质, 我们对本书的翻译工作进行了精心的组织. 首先是北京大学金融数学系 2002 级一批优秀的本科生和硕士生同学帮助我完成第一稿的基础性翻译工作, 然后在他们之间进行交换审阅, 再由我本人总校一遍, 最后又请金融数学系的研究生张松、马晓静、陈琴、方圆、王璐璐和李欣进行第二审, 最终由我全篇通审. 参与翻译工作的学生还有: 陈肖安、周清、叶明、张志强、郑江平、李冬来、崔庸非、朱楠、方家聪和李凌飞.

正是由于他们的辛勤工作, 才使得我在全书总校工作中做得很顺利. 在此向所有参加本书翻译工作的学生表示感谢! 另外, 在本书翻译进入最后阶段时, 我有幸参加了国家科技部 973 项目《金融风险控制中的定量分析与计算》(编号 2007CB814900), 作为课题《银行与保险业中的风险模型与数据分析》(编号 2007CB814905) 的课题负责人, 因此本书的翻译工作也得到该项目的一定资助, 在此表示感谢. 同时, 我也希望本书的翻译工作能对该课题的开展起到很好的推动作用.

当然, 对于这样一本精算技术的专业著作, 我们的理解难免有不足之处, 译文中不妥之处恳请读者指出.

译 者

2008 年 7 月于北京大学

前　　言

在本书第 1 版的前言中, 我们这样解释了写作本教材的目的.

本教材是围绕如下的一个基本原则进行组织的：精算科学的主要内容是构造和分析数学模型, 这些模型刻画了资金流入和流出保险系统的过程. 而对整体系统的全面分析超出了任一本教科书的范围, 所以我们将主要关注损失过程, 也就是说因保险赔付而产生的现金流出.

我们并不要求读者已经具有很好的保险系统的知识背景. 本书中的保险术语在首次出现时都会给出其定义. 实际上, 本教材的大部分内容可以脱离开保险这个背景而重新组合在一起. 同时, 本书又是一本统计应用的教材. 我们尽可能使本书的例子集中于保险背景, 用保险行业的语言和资料组织我们的材料, 并着力避免涉及那些精算实践中很少使用的统计方法.

特别地, 本教材在 1998 年的第 1 版达到了以下 3 个目的.

(1) 对 1984 年出版的由 Robert Hogg 和 Stuart Klugman 所著的《损失分布》[59] 中的分布拟合的内容进行了更新.

(2) 对 1992 年出版的由 Harry Panjer 和 Gordon Willmot 所著的《保险风险模型》[106] 中离散分布和聚合风险模型计算的内容进行了更新.

(3) 将北美精算协会的强化讨论班 152(应用风险理论) 中 3 个作者的材料进行了综合整理.

本书第 1 版出版后不久, 北美产险精算师协会 (Casualty Actuarial Society) 和北美精算师协会就修改了其考试大纲, 将本书第 1 版列为考试参考书. 令人高兴的是, 本书第 1 版被选为新课程体系课程 3 和课程 4 的主要参考材料. 遗憾的是, 本书的主要素材被两门课程分割了, 而其分割方式并不符合本书最初的组织安排. 于是我们有充分的理由来修订第 1 版. 此外, 还有其他的考虑.

(1) 第 1 版假设读者很熟悉数理统计的内容, 这也是写作本书时精算考试的内容, 但是随后被逐渐淡化. 一些关于数理统计的背景材料在本版的第 9 章给出.

(2) 长期以来, 精算考试都包括生存模型的内容, 这个模型是用于确定身故、失效或伤残时间的概率模型, 它与确定索赔量或索赔数的一般概率模型并没有本质的差异. 因此, 本版将这些内容整合在一起, 更强调建立实证性模型. 第 10 章和第 11 章将研究这部分内容.

(3) 最近几年的精算考试大纲所去掉的内容中, 下面两项内容是应该再加上的, 至少应该将其主要部分保留. 一项内容是修匀方法, 是数值观测序列进行平滑和插值处理的方法. 第 15 章将讨论这部分内容. 另一项内容是对初步的估计公式的调整, 例如在生命表数据的研究中要处理海量的数据时就需要这种调整. 11.4 节将讨

论这部分内容.

(4) 第 1 版扼要地论述了随机模拟的内容, 第 17 章将对这部分内容进行一定的扩充.

对于一些继续保留的内容, 除了重新安排模型本身和建模方法外, 最本质的改变是重写了破产理论部分 (第 7 章和第 8 章), 并且对有限波动信度公式给出了更好的解释 (第 16 章).

同时, 我们努力将所有素材整合在一个具有逻辑关系的单独精算模型的构造过程中, 并同时保持各部分的相对独立性, 因此, 本书划分为 5 个部分.

自从本书第 1 版出版以来, 人类的计算能力在不断提高. 第 1 版提供了计算最大似然估计和总损失的 DOS 程序, 这些程序仍然可在 Wiley 出版社的网站上获取: ftp://ftp.wiley.com/public/sci_tech_med/loss_models/.

此外, 例子和习题中使用的数据也放在相应的文件中. 但是, 读者很可能会采用 Microsoft Excel® [①] 之类的制表程序进行计算. 在本书的很多地方我们将给出 Excel® 的指令或命令, 这并不意味着作者对该软件有任何的偏好和支持, 我们只是在借用这个工具软件进行说明.

和第 1 版一样, 本版的许多习题取自北美产险精算师协会和北美精算师协会精算考试试题, 但是用本书的记号和术语重新表示, 并去掉了 5 个选择答案. 这类习题用 (*) 标示. 当然，这些问题并不对这个考试今后的题目有任何的代表意义.

S. A. Klugman　H. H. Panjer　G. E. Willmot
美国爱荷华州得梅因市
加拿大安大略省滑铁卢市

① Microsoft® 和 Excel® 表示微软公司在美国或其他国家的注册商标或商标.

致 谢

Stuart Klugman:

虽然编写本版不像第 1 版那样是一个令人感到筋疲力尽的任务, 但是仍然有许多工作要做并且也是在很多人的帮助下才完成了本版的修订工作. 有很多人对改进第 1 版提供了修正和改进的建议, 尤其是 Elias Shiu 和 Don Minassian 全面地指出了本书的不足之处. Clive Keatinge 也在正式出版前提出了许多修改意见. 当北美精算协会决定需要修订第 1 版以适应其作为两门考试课程的参考资料时, 在 Clive Keatinge 的领导下成立了一个委员会. 这个委员会制定的计划体现在本书的第 2 章至第 5 章和第 9 章至第 13 章. 我们也在这些章节依赖于这个委员会的专业性工作增加了一些内容以使其更全面.

特别要感谢我的合作者及时完成了相应部分的工作.

在修改这个致谢时, 正值我的妻子 Marie 在长期与多种疾病斗争中不幸去世后不久. 不管是在她健康时还是生病期间, 她都一直在鼓舞我, 也在鼓舞所有认识她的人们. 最后的初稿能够按时交付出版社, 不仅因为我对这项工作的热爱, 更多的是受她的影响, 因为遵守承诺和按约定行事是她的价值观.

Harry Panjer:

修订或出版图书决不是一件好玩的事情. 但是, 在本版的修订过程中, 我们确实在增加大量新材料的过程中, 感受到了很多乐趣. 每个作者在全书的新材料中各有贡献. 我有幸编写了样条插值和平滑这一章的初稿, 它为传统的精算修匀问题提供了一种现代的处理. 我非常感谢 Stuart Klugman 和 Drake 大学的学生们, 他们全面检查了这些材料并发现了一些错误 (因此避免了我在将来可能的尴尬).

Stuart Klugman 是本项目有远见的、精神上和实践中的领导. 我也感谢 Gordon Willmot, 他在加拿大滑铁卢大学领导了一支进行证明推导工作的团队. 本版如果还有错误, 相对第 1 版来说一定是微不足道的. 他们确实做了很多工作. 最后, 感谢我的妻子 Joanne Coyle, 她容忍了我将很多周末和夜晚消磨在办公室中. 我们的儿子 Lucas D. 和 Lucas R., 见证了本书第 1 版的写作过程, 而现在他们已经长大并离开了家. 只有当有人告诉他们本书的致谢中提到他们时, 他们才会知道出版第 2 版的事情.

Gordon Willmot:

本版的写作过程经历了相当多不断重新组织和更新的工作. 这些工作反映了本书在对一些内容进行组合时的创新性. 我们尝试提供一种一体化的方法来利用数据建立模型, 并进而利用这些模型来进行定价或其他应用. 严格说来, 最终的成文不容置疑地得到了无数人的帮助, 人数太多以至无法历数. 但是, 还是需要特别感

谢以下各位对本书第 2 版的无价帮助: Catherine Donnelly, Steve Drekic, Mary Lou Dufton, Jessica Ling-Wai Lam 和 Claire Xiao-Dan Yang.

我还要对合作者们表示诚挚的谢意, 特别是 Stuart Klugman 对本项目的先锋作用, 带领我们沿着正确的路线前进, 并在处理各种情况 (特别是很困难的情况) 时是他表现出了近乎无穷的耐心. 我也要再次对我的妻子 Deborah 以及女儿 Rachel, Lauren 和 Kristen 表示感谢, 有了她们内在和外在的牺牲才使得我能够有时间和精力来完成这项工作.

S.A.K., H.H.P., G.E.W.

目　录

第四部分　参数化统计方法

第一部分　引　　言

第1章　建　　模

1.1　模型化方法

在对任何已知的问题进行研究时都可以考虑模型化方法. 精算学中的很多问题都涉及建立数学模型, 这种模型可以用于预测或预计未来的保险成本.

模型是一种对现实简化的数学表达, 精算师根据自己的知识和经验并基于历史数据构造所需的模型. 实际数据对精算师的指导作用不仅体现在模型形式的适应性, 还会帮助精算师校准一些未知量, 通常称这些未知量为**参数**. 最终的模型是对简洁性和可得数据的适应性两方面平衡的结果.

我们可以通过一些标准来度量模型是否简单, 例如未知参数的个数 (越少越简单); 我们还可以通过分析数据与模型的差异程度来度量模型对数据的适应性. 最终所选择的模型是对两个方面权衡的结果, 即模型的拟合程度和简单程度.

1.1.1　建模流程

图 1-1 是建模流程的示意图, 这个流程由以下 6 个阶段组成.

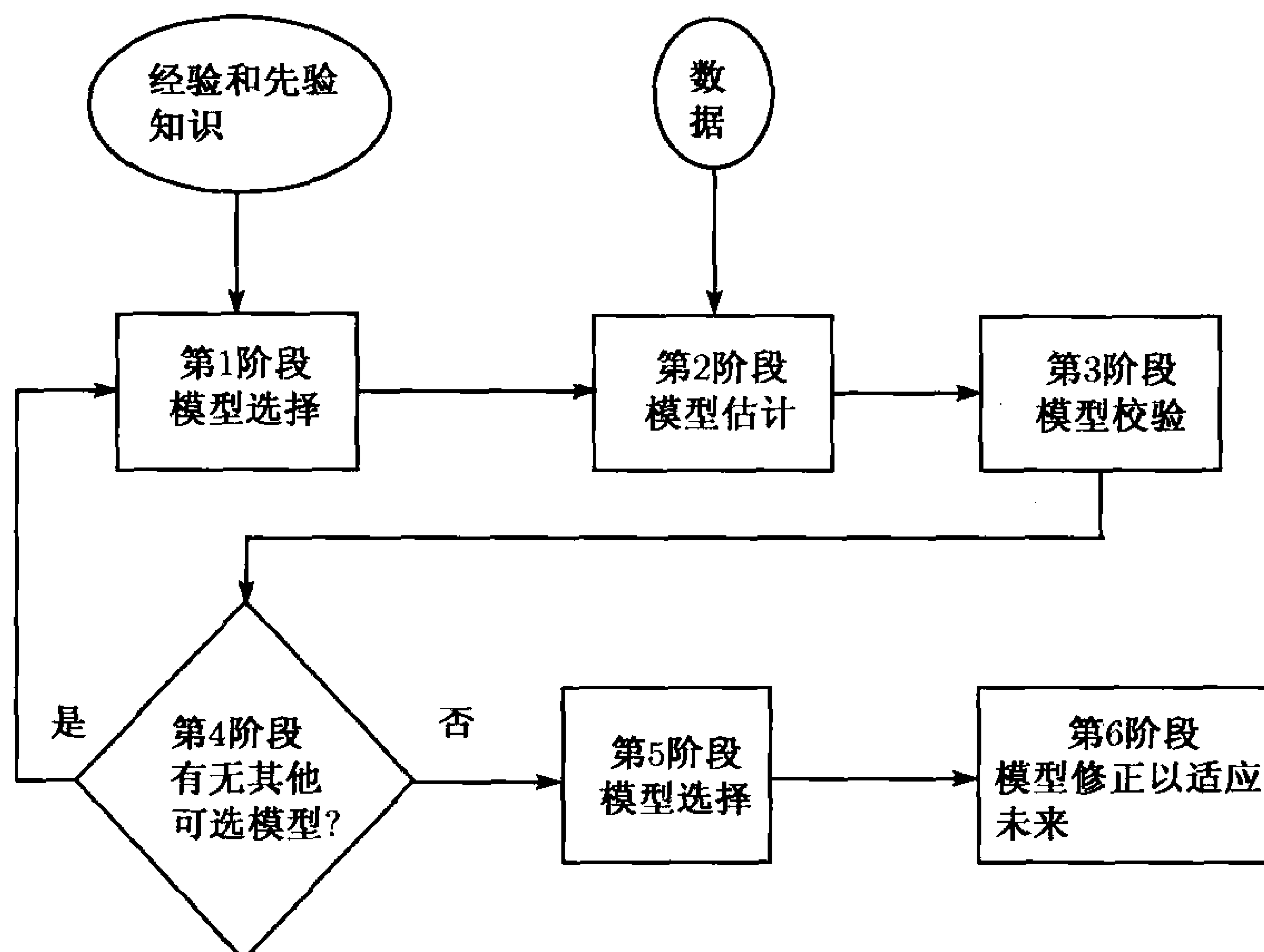


图 1-1　建模流程

第 1 阶段 根据分析人员对现有数据的性质和形式的先验认知和经验, 初步选择一个或多个模型. 例如, 在研究死亡率时, 所选的模型也许会包含以下这些协变信息: 年龄、性别、保单生效期限、保单类型、健康方面的信息和生活方式等变量. 在研究保险的损失量大小时, 会对统计分布类型 (例如, 对数正态、gamma、Weibull) 有一些自然的选择.

第 2 阶段 基于观测数据进行模型的校准. 在研究死亡率时, 数据可能是某寿险保单群体的信息. 在研究财产保险的索赔时, 数据可能是某财产保险群体的实际赔付数据.

第 3 阶段 确定所拟合模型的有效性, 依赖于它是否充分考虑了数据中的所有信息. 这里可以采用各种诊断检验. 例如一些著名的统计检验: 卡方拟合优度检验、Kolmogorov-Smirnov 检验, 或者按照事物的本质进行的定性检验. 检验方法的选择直接依赖于建模的目的. 在与保险有关的研究中, 经常要求拟合的模型能够从整体上复制出实际经验数据所表现的损失. 在保险实务中也常称之为模型的**无偏性**.

第 4 阶段 要有适当的机会考虑其他可选模型. 特别是, 如果在第 3 阶段揭示出前面的模型不适用, 这个步骤将特别有价值. 在这个阶段也许会考虑不只一个有效模型.

第 5 阶段 将所有第 1 阶段至第 4 阶段考虑的有效模型按照一定的准则进行比较选择. 这时可以利用前面获得的一些检验结果来选择, 也可以考虑其他的准则. 一旦某个模型胜出, 则淘汰的那些模型可用于敏感性分析.

第 6 阶段 最终要保证被选出的模型适用于未来的应用. 也许要对参数进行适当的调整, 以反映从观测数据的时期到未来模型使用时期之间的可预料的通货膨胀变化.

当新的数据产生后或者环境有所变化时, 需要重复进行以上 6 个步骤以改进模型.

1.1.2 建模方法的优势

在解释模型方法的优势时, 需要将模型方法与其他严格基于经验进行决策的方法进行比较. 基于经验的方法假设未来将与由已发生的现象抽取的样本具有完全相同的表现, 也许只需要对例如通货膨胀之类的基本要素进行调整. 下面的例子将对此予以说明.

例 1.1 *某团体人寿保险合同由不同年龄和受益水平的 1 000 个雇员组成. 在过去的 5 年中, 已有 14 名雇员身故, 共得到 580 000 元 (由于该计划的身故受益与雇员的工资水平挂钩, 所以需要进行通货膨胀调整) 的赔付. 试根据以上信息对该合*

同下一年的预期身故赔付进行经验估计.

解 下一年预期身故赔付的经验估计为 116 000 元 (5 年赔付总额的 1/5), 考虑到受益水平增加的因素, 这个数值还会有所增加. 当然, 这样估计的危险在于, 过去 5 年的经验不一定完全能够反映这个合同在未来一年的情况, 因为在如此短的时期内身故赔付的表现可能会有很大波动. □

看来更合理的方法是建立一个模型, 上面例子的情形就是生命表. 要构造这样的表需要积累很多个体的经验, 上面例子中 1 000 个人的经验是不够的. 有了这张生命表, 不仅可以估计下一年的预期赔付, 还可以度量我们所作的估计本身的风险, 这种风险一般是通过计算下一年赔付的标准差或赔付分布函数的某个分位点进行度量的. 这也是《精算数学》([16]) 用了很大篇幅进行详细论述的问题.

这种方法由北美精算协会负责研究精算基本原理的专业委员会以专业的方式进行了规范. 公开发表的"精算学基本原理"([123]) 第 571 页的原理 3.1 指出"精算风险可以用随机模型的方法进行表述, 这些模型是基于对这些精算风险变量未来的概率分布以及未来的环境状况的假设而建立的". 这里的精算风险变量一般指: 风险是否发生、发生的时间和损失量, 即索赔事件的发生机率、如果索赔事件发生其发生的时间以及解决索赔所需的成本.

1.2 本书的结构

本书将带领读者领略整个建模过程, 但并不是按照前面介绍的几个阶段的顺序来讲解. 最好的应用模型与最好的学习模型是有所不同的. 在本书中, 我们首先要研究模型本身并介绍如何使用这些模型, 然后讲授如何确定该使用哪个模型. 这样做的理由是, 你无法凭空选择适用的模型, 只有当分析人员对于一组可选的模型具有全面的认识后, 才能将其选择的范围缩小. 基于这种考虑, 本书采用了如下所示的结构.

(1) 概率基础. 显然, 不确定 (未定) 事件中必然隐含着概率模型. 第 2 章和第 3 章将复习随机变量和其他可能与模型有关的基本计算, 包括矩和分位点计算.

(2) 充分理解概率分布. 为了选择适用的概率模型, 分析师应该掌握相当多的概率模型. 此外, 为了对模型有一个好的先验的选择 (有时为了成本等方面的考虑, 必须有一些先验的选择), 还必须充分理解这些模型的特点. 第 4 章介绍了各种分布模型以及它们的主要特征, 这里包括离散分布和连续分布.

(3) 保险赔付的调整. 一般情况下, 保险合同不可能对标的损失提供全额的赔偿. 例如, 合同中常常含有免赔 (例如, 合同只赔偿损失超过 250 元的部分)、最大限额 (例如, 对于任何一次损失事件, 合同最多赔偿 10 000 元) 等保险责任条款. 这种调整将影响最终的损失概率分布以及相关的计算 (例如各阶矩). 第 5 章将具体

说明如何进行这些调整.

(4) 总损失和破产. 至此, 模型不仅要反映每次赔付的金额还要反映赔付的次数. 在对一些保单组合、某个业务线或者某个公司进行损失建模时, 我们将关心赔付的整体情况. 当模型兼顾关于赔付次数和每次赔付金额的概率时, 称之为总损失模型, 第 6 章将介绍这类模型的计算. 通常情况下, 赔付是按时间顺序发生的. 当赔付金额很大时, 有可能模型所考虑的载体 (例如保险公司) 在某个时刻将无法进行正常的支付, 一般称这种状态为破产. 第 7 章和第 8 章将分别建立计算破产事件发生概率的模型.

(5) 数理统计基础. 因为考虑的大多数模型是概率模型, 所以需要数理统计方法进行估计并决策. 第 9 章并没有全面重复一般数理统计课程或教材的内容, 只是给出了本书后面部分所需要的一些必要的知识.

(6) 经验模型的构造. 有时人们需要利用数据的经验分布, 这也许是因为数据的规模足够大或者是因为需要很好地表现数据本身的特点. 第 10 章和第 11 章将主要讨论这部分内容, 包括对原始数据直接计算的简单情形, 对截断或删失数据的调整, 以及对大数据集特别是那些在生存模型研究中出现的数据的适当修正.

(7) 建立参数化模型. 通常情况下, 对数据进行适当的平滑处理, 然后提出一个总体分布, 是非常有意义的. 第 12 章提出了基于前面的模型假设的参数估计方法. 第 13 章将考虑模型选择问题.

(8) 第 14 章汇总了一些实例, 总结和集中表现目前为止我们已经介绍的内容.

(9) 估计的调整. 有时, 需要对结果做进一步的调整. 本书介绍了两种调整方法. 首先是插值和平滑 (也称为修匀) 处理, 第 15 章介绍了这部分内容, 主要是三次样条方法. 有时没有简单适用的已知概率分布, 例如死亡时间, 而经验分布可能不像已知总体分布那样光滑, 所以, 需要采用修匀方法进行必要的调整. 其次, 还有一种情况是几个估计都是基于很小的观测量得到的, 这时可以考虑添加一些相关的观测来提高估计的精度, 但是必须注意这些观测是否来自不同的总体. 第 16 章介绍的可信度理论考虑了增加观测时如何进行适当调整的机制和方法.

(10) 随机模拟. 在很难得到解析结果时, 随机模拟 (利用随机数) 方法也许可以提供一些答案. 第 17 章简单介绍了这种技术.

第二部分　精 算 模 型

第2章 随机变量

2.1 引 言

一般的精算模型将尝试表现未来不确定的支付流. 这种不确定性包括: 任何事件或所有事件 (某种支付) 是否发生、发生的时间 (何时进行支付?) 和损失量 (支付了多少?). 因为表现不确定性的最常用的方法是概率, 所以我们也将重点考虑概率模型. 在所有的情形中, 都假定相关的概率分布已知. 我们将在第 11 章至第 13 章中讨论如何选择适用的分布. 本章将考虑如下与精算概率模型有关的内容.

(1) 随机变量的定义及其相关的重要函数和一些例子.

(2) 围绕概率模型的基本计算.

(3) 具体的概率分布以及性质.

(4) 利用索赔程度模型进行更高级的计算.

(5) 考虑随机支付的个数也是随机的概率模型.

(6) 描述公司盈余过程的模型.

有两个重要的模型本书没有考虑. 第一个模型是关于未来投资收入的模型. 虽然这些模型方法也属于本书的范围, 但是产生这些模型的金融背景超出了本书的主要写作目的. 第二个模型是考虑支付发生时刻的利息收入模型 (随机的或确定的), 这类模型的简单情形可能会在一些例子中涉及, 但是如果要全面彻底地介绍这类模型在寿险和年金险中的应用则另外需要一本教材, 例如《精算数学》[16].

这里所寻求的共性是所有表现随机现象的模型都具有的共同的内在因素, 每个这种模型的结果都是一个集合, 而其中一些特殊结果的发生将决定企业经营的成功. 各种可能结果所附着的概率将帮助我们定量地刻画预期结果以及不出现这些结果的风险. 基于这样的考虑, 一般总是将这些隐含的随机变量记为 X 或 Y, 但是, 在本书的各个具体问题中也会为这些变量命名并提供某些具体性质. 当然, 也有一些精算模型并不是这样的, 例如, 人寿保险中的**办公系统模型**由如下元素组成: 保单类型、年龄范围、性别等.

为了推广这个概念, 这里给出关于"精算科学的基本原理"① 的最近一个工作草案中的定义.

---

① 这个文件是北美财产险精算协会 (CAS) 和北美精算协会的联合工作组正在进行的一个项目的工作报告, 关键的原则是: 目前存在的模型应该代表某种精算现象, 并且有充分的数据用于模型的校准.

现象是指那些可以观测到的发生. **试验**是在一定条件下对某给定现象的一个观测. 一次试验的最终观测称为**结果**. 事件是一个或多个可能结果的集合. 随机现象是指其试验可能会有一个以上的可能结果. 具有**随机现象**的事件称为**不确定结果**. **概率**是对一个事件的各种结果发生可能性的度量, 这个度量将这种可能性进行了标准化处理, 按照从 0 增加到 1 的数值表示. **随机变量**是一个函数, 它对每个可能的结果赋予一个数值.

下面列出了在精算工作中涉及的各种随机变量.

(1) 随机选择的某个生存个体的死亡年龄 (**模型 1**);

(2) 对于随机选择的某个个体, 从购买保险到死亡发生的时间间隔;

(3) 对于随机选择的某个工伤事故的索赔者, 从其伤残发生至康复或死亡的时间间隔;

(4) 对于随机选择的某个索赔, 从其事故发生到报告给保险人的时间间隔;

(5) 对于随机选择的某个索赔, 从其报告给保险人到赔付完全结束的时间间隔;

(6) 对于随机选择的某个寿险索赔, 实际支付的理赔金额;

(7) 对于随机选择的某个机动车辆车身损失索赔, 实际支付的理赔金额 (**模型 2**);

(8) 对于随机选择的承保机动车辆, 在一年内发生车身损失索赔的次数 (**模型 3**);

(9) 对于随机选择的某个医院, 在一年中发生医疗事故索赔的总金额 (**模型 4**);

(10) 对于随机选择的住房抵押贷款保险提前中止的某个承保个体, 其违约或提前还款前经过的时间;

(11) 对于随机选择的高收益债券, 满期时的给付金额.

由于所有这些现象都可以用随机变量表示, 概率和数理统计的机制可以使得我们对这些现象建立模型并进行分析. 2.2 节讨论在描述一个随机变量的过程中所使用的 5 个关键函数. 它们将用到上面介绍的并将不断使用的 4 个模型以及另外 2 个模型进行说明.

2.2　重要函数和 4 个模型

定义 2.1　*对于某个随机变量X, 累积分布函数也称作分布函数表示X小于或等于某个给定数值的概率, 通常记为$F_X(x)$或$F(x)$*[①]*. 即: $F_X(x) = \Pr(X \leqslant x)$, 也常用简写cdf表示这个函数.*

① 当讨论某个随机变量的某种函数时, 通常用脚标来代表该随机变量, 而且需要对不同的随机变量进行区别时才这样使用. 此外, 在引入 6 个模型时进行了简化, 例如随机变量 2 的分布函数不是表示为 $F_{X_2}(x)$, 而是简单记为 $F_2(x)$.

分布函数必须满足以下的必要条件①.

- 对所有 x, $0 \leqslant F(x) \leqslant 1$.
- $F(x)$ 是非降的.
- $F(x)$ 是右连续的②.
- $\lim_{x\to-\infty} F(x) = 0$且 $\lim_{x\to\infty} F(x) = 1$.

由于不必要求左连续, 所以分布函数可能是带跳的, 在跳跃点的函数值取跳上去的高点.

前面讨论的 4 个模型的分布函数如下.

模型 1 这个随机变量可以用来刻画死亡年龄的模型, 所有介于 0 和 100 之间的年龄都可以用这个变量来刻画. 尽管现实中人的寿命都是有限的, 但是也可以采用没有上界的模型, 只是在极其高的年龄点赋予很低的概率, 这样就可以避免在建模时必须规定一个寿命上限的问题.

$$F_1(x) = \begin{cases} 0, & x < 0, \\ 0.01x, & 0 \leqslant x < 100, \\ 1, & x \geqslant 100. \end{cases}$$ □

模型 2 这个随机变量可以作为机动车辆保险赔付金额的度量, 可能取所有的正值. 与死亡数据相同, 这里也应该考虑一个上限 (世界上所有的货币都来自人的想象), 但是, 这个模型说明当建模时进行了现实性考虑后则是不完美的模型.

$$F_2(x) = \begin{cases} 0, & x < 0, \\ 1 - \left(\dfrac{2\ 000}{x + 2\ 000}\right)^3, & x \geqslant 0. \end{cases}$$ □

例 2.2 试绘出模型 1 和模型 2 分布函数的图形 (其他模型要求在习题 2.2 中给出其图形).

解 见图 2-1 和图 2-2. □

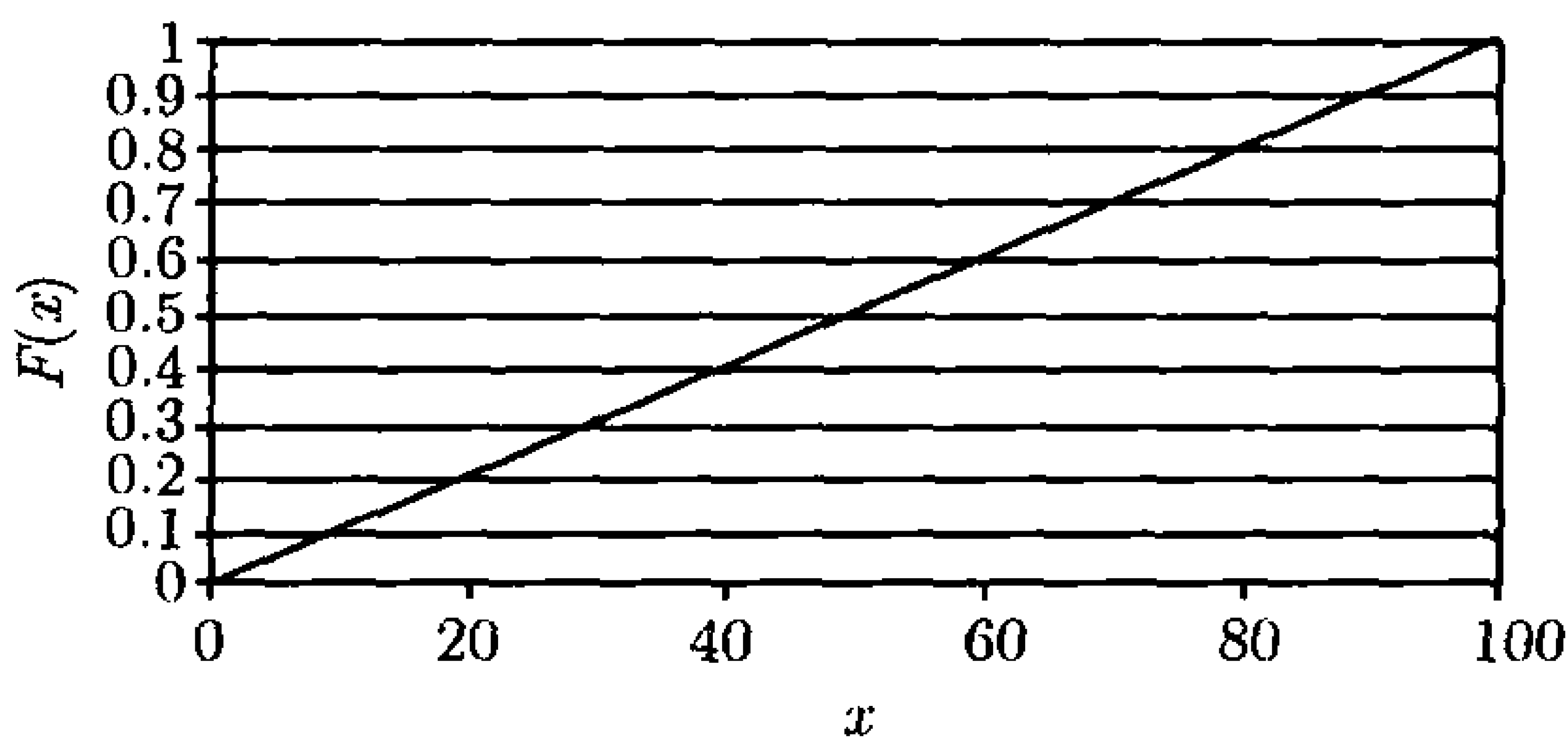


图 2-1 模型 1 的分布函数

① 第一条是后面三个条件的自然结果.

② 右连续表示对于任意点 x_0, 当 x 从 x_0 的右边趋向于 x_0 时, $F(x)$ 的极限值为 $F(x_0)$. 这个结论对 x 从 x_0 的左边趋向于 x_0 时不一定成立.

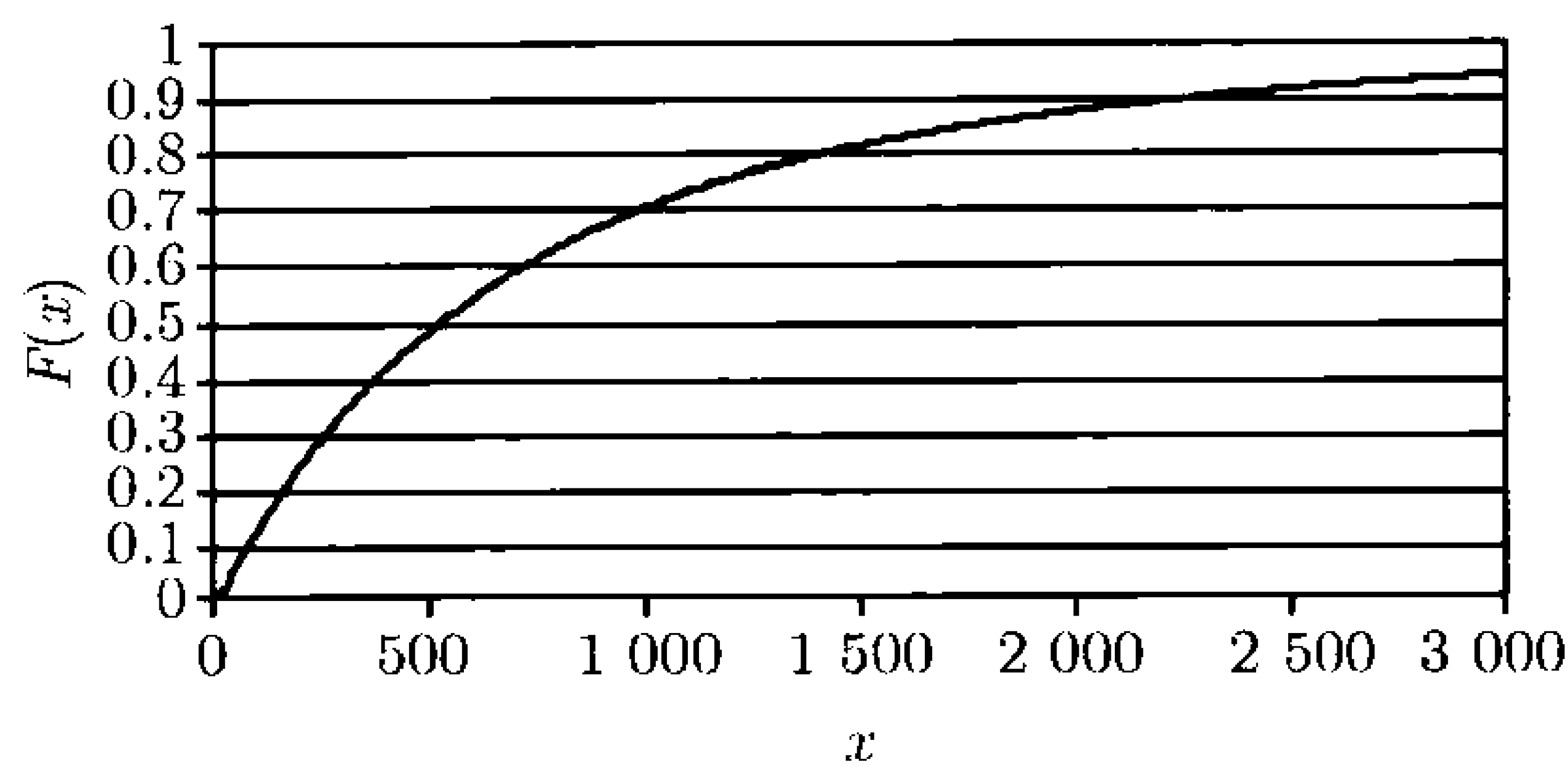


图 2-2 模型 2 的分布函数

模型 3 这个随机变量可以用来刻画每个保单的年度索赔次数的模型, 这时概率集中于 (0,1,2,3,4) 这 5 个点, 而且每个点的概率就是分布函数在该点的跳跃幅度. 虽然这个模型对索赔数有最大值的限制, 但也可以用无上限的分布 (例如 Poisson 分布) 进行建模.

$$F_3(x)=\begin{cases}0, & x<0,\\ 0.5, & 0\leqslant x<1,\\ 0.75, & 1\leqslant x<2,\\ 0.87, & 2\leqslant x<3,\\ 0.95, & 3\leqslant x<4,\\ 1, & x\geqslant 4.\end{cases}$$ □

模型 4 这个随机变量可以用来刻画医疗事故保险每个保单的年总索赔额模型, 因为在很多年份没有索赔发生, 所以这个模型的主要概率 (0.7) 集中在零点, 剩下的概率 0.3 分布在正实数中.

$$F_4(x)=\begin{cases}0, & x<0,\\ 1-0.3\mathrm{e}^{-0.000\,01x}, & x\geqslant 0.\end{cases}$$ □

定义 2.3 随机变量的支集是由该随机变量所有可能的数值组成的集合.

定义 2.4 若随机变量的支集最多包含可数个值, 则称之为离散型随机变量. 若随机变量的分布函数是连续的而且除可数个点外是处处可微的, 则称之为连续型随机变量. 若随机变量不是离散型, 而且除至少一个点、至多可数个点外是处处连续的, 则称之为混合型随机变量.

上面定义的三种随机变量不可能穷尽所有的随机变量类型, 但是对于本书的讨论足够了. 离散型随机变量的分布函数除在具有正概率的跳跃点外均为常数. 而混合型随机变量分布函数则至少有一个跳跃点. 连续型随机变量的可微性要求是为了使其除可数个点外密度函数 (后面将给出定义) 存在.

例 2.5 对于前面的 4 个模型, 确定每个模型的支集和随机变量的类型.

解 模型 1 的分布函数是连续的, 而且除了 0 和 100 这两个点外均可微, 所以是连续型分布, 支集为 0 到 100 之间的数, 但不是很确定支集是否包含 0 和 100 这两个点. 模型 2 的分布函数是连续的, 而且除 0 点外均可微, 所以是连续型分布, 支集为除 0 点之外的所有正实数. 模型 3 的随机变量只是在点 0,1,2,3 和 4 有概率, 所以是离散型分布. 模型 4 的分布函数除了在点 0 有跳跃外分布函数是连续的, 支集为非负实数, 这是一种混合分布. □

这 4 个模型代表了分布函数最常见的方式. 在本书随后的部分中, 对于一些与分布函数类似的函数我们只讨论其在随机变量支集范围内的取值.

定义 2.6 *随机变量X的生存函数通常记为$S_X(x)$或$S(x)$, 表示X大于某个给定值x的概率, 即$S_X(x)=\Pr(X>x)=1-F_X(x)$.*

由上面的定义, 自然有:

- 对所有的 x, $0\leqslant S(x)\leqslant 1$;
- $S(x)$*是不增的*;
- $S(x)$*是右连续的*;
- $\lim\limits_{x\to-\infty}S(x)=1$*且* $\lim\limits_{x\to\infty}S(x)=0$.

由于生存函数不一定是左连续的, 所以可能会出现向下的跳跃点, 在这些点, 函数值取其跳跃点的下 (右) 端点.

因为生存函数是分布函数的补函数, 所以, 当其中的一个已知时自然可以推出另外一个. 生存函数最初的使用是为了刻画时间变量, 当随机变量刻画金额变量时通常采用分布函数.

例 2.7 *给出前面 4 个模型的生存函数.*

解

$$S_1(x)=1-0.01x,\qquad 0\leqslant x<100,$$

$$S_2(x)=\left(\frac{2\ 000}{x+2\ 000}\right)^3,\quad x\geqslant 0,$$

$$S_3(x)=\begin{cases}0.5, & 0\leqslant x<1,\\ 0.25, & 1\leqslant x<2,\\ 0.13, & 2\leqslant x<3,\\ 0.05, & 3\leqslant x<4,\\ 0, & x\geqslant 4,\end{cases}$$

$$S_4(x)=0.3\mathrm{e}^{-0.000\ 01x},\quad x\geqslant 0.$$

□

例 2.8 *给出模型 1 和模型 2 的生存函数的图形.*

解 见图 2-3 和图 2-4. □

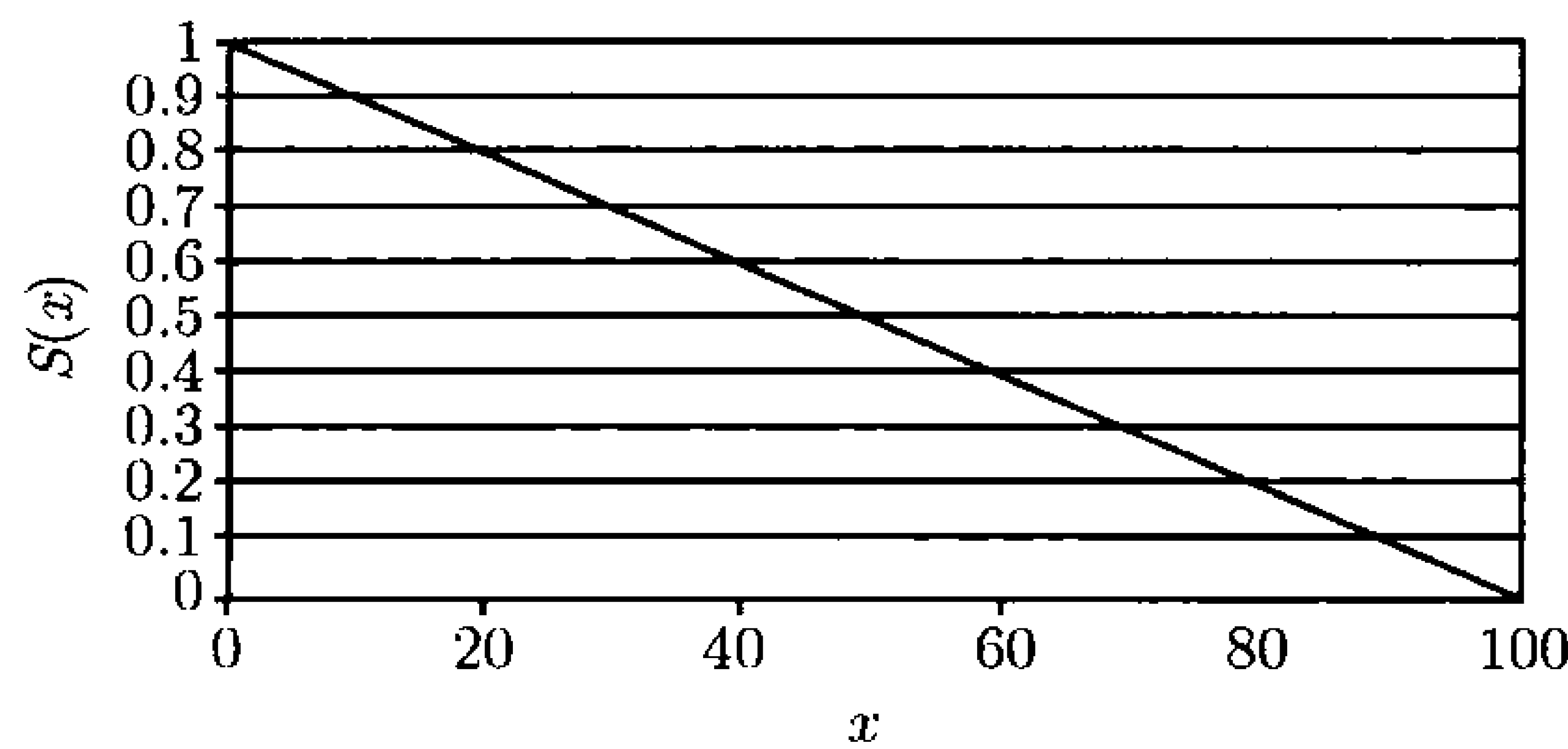


图 2-3 模型 1 的生存函数

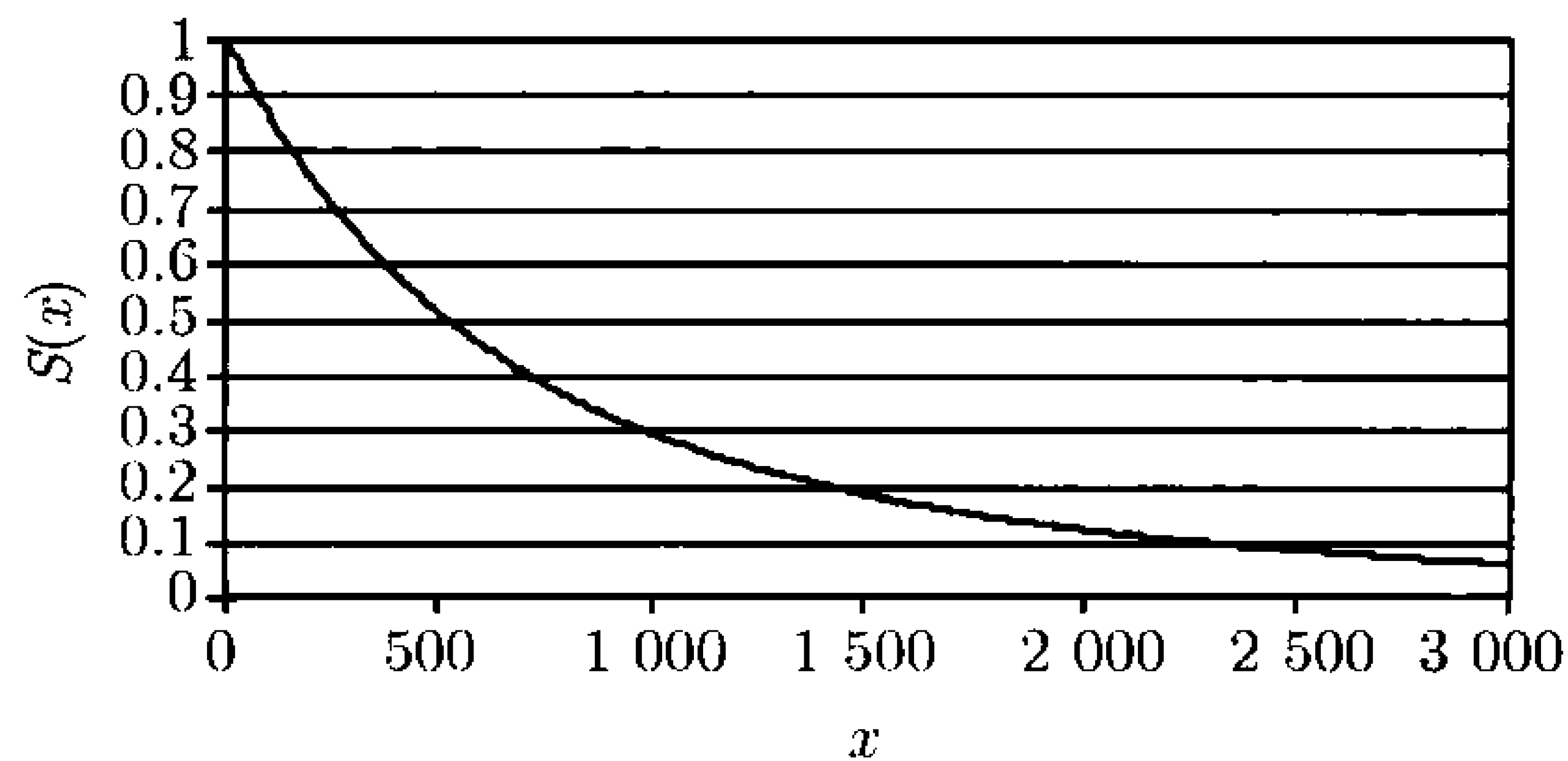


图 2-4 模型 2 的生存函数

无论是分布函数还是生存函数都可以用来确定概率值. 令 $F(b-) = \lim_{x \to b^-} F(x)$, 类似地可以定义 $S(b-)$, 这表示 x 从 b 的左边趋向 b 时的极限. 因此有: 对任何属于 X 支集的两个实数 $a < b$[①], 有

$$\Pr(a < X \leqslant b) = F(b) - F(a) = S(a) - S(b),$$

$$\Pr(X = b) = F(b) - F(b-) = S(b-) - S(b).$$

当分布函数在 x 点连续时, $\Pr(X = x) = 0$; 否则, 概率值就是跳的幅度. 下面定义的两个函数与概率的关系更为直接, 分别对连续和离散分布给出定义.

定义 2.9 随机变量X的概率密度函数, 简称为密度函数通常记为$f_X(x)$或$f(x)$, 它表示分布函数的导数或生存函数导数的负值. 即$f(x) = F'(x) = -S'(x)$, 密度函数只在分布函数导数存在的点上有定义, 有时也采用*pdf*缩写表示随机变量的概率密度函数.

虽然密度函数不能直接给出概率值, 但还是可以提供相关的信息. 随机变量在密度函数取值较高的区域发生的可能性将高于密度函数取值较低的区域. 对于给定区间的概率值、分布函数值、生存函数值都可以通过对密度函数的积分得到. 即如果密度函数在所考虑的区间内有定义, 则有 $\Pr(a < X \leqslant b) = \int_a^b f(x)\mathrm{d}x$, $F(b) =$

① 此处为译者加入的说明. —— 译者注

$\int_{-\infty}^{b} f(x)\mathrm{d}x$ 和 $S(b)=\int_{b}^{\infty} f(x)\mathrm{d}x$.

例 2.10 给出前面 4 个模型的概率密度函数.

解

$$
\begin{aligned}
&f_1(x)=0.01, \quad 0<x<100,\\
&f_2(x)=\frac{3(2\ 000)^3}{(x+2\ 000)^4}, \quad x>0,\\
&f_3(x)\quad \text{未定义},\\
&f_4(x)=0.000\ 003\mathrm{e}^{-0.000\ 01x}, \quad x>0.
\end{aligned}
$$

应当注意的是模型 4 的密度函数不能完全描述概率分布函数, 作为一个混合分布, 在 0 点也有离散的概率. □

例 2.11 给出模型 1 和模型 2 的密度函数的图形.

解 见图 2-5 和图 2-6. □

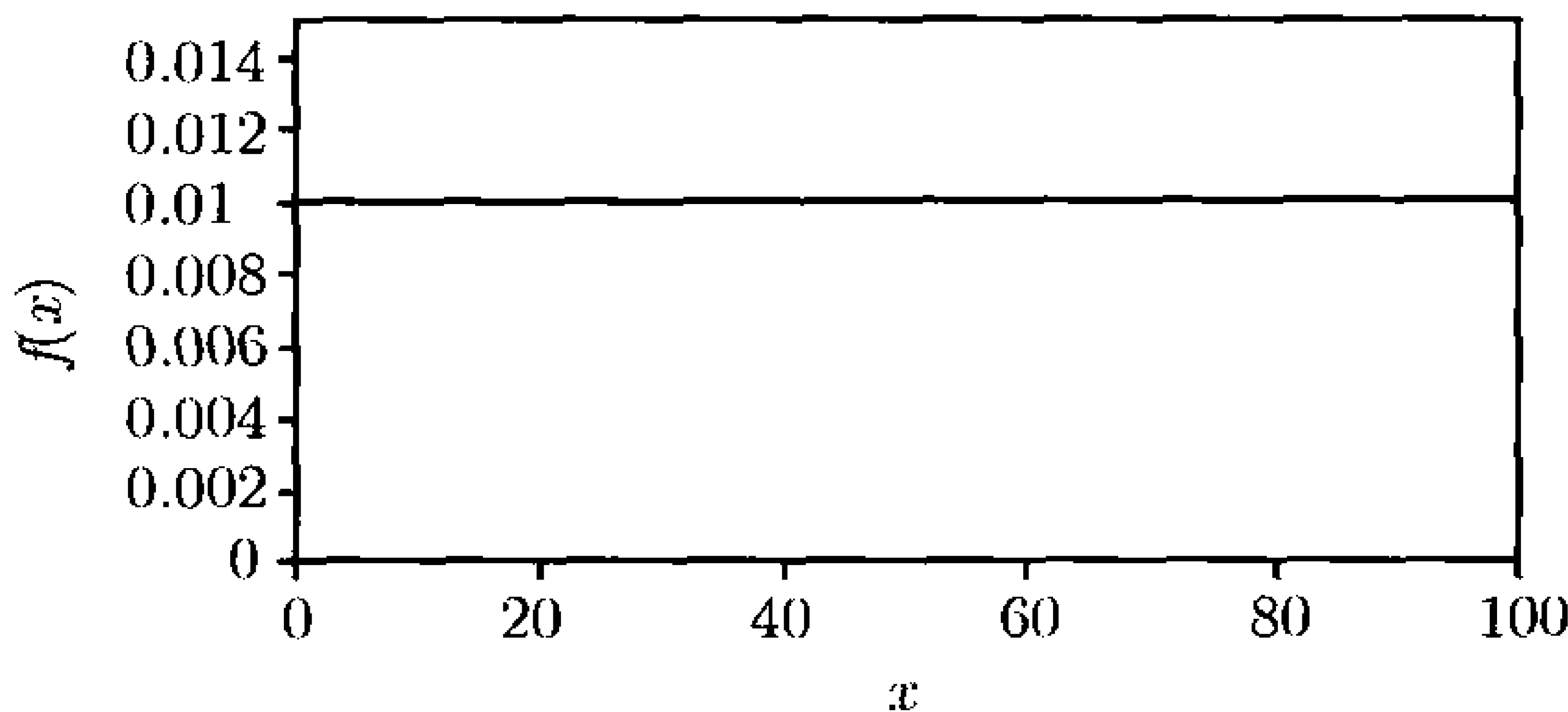


图 2-5 模型 1 的密度函数

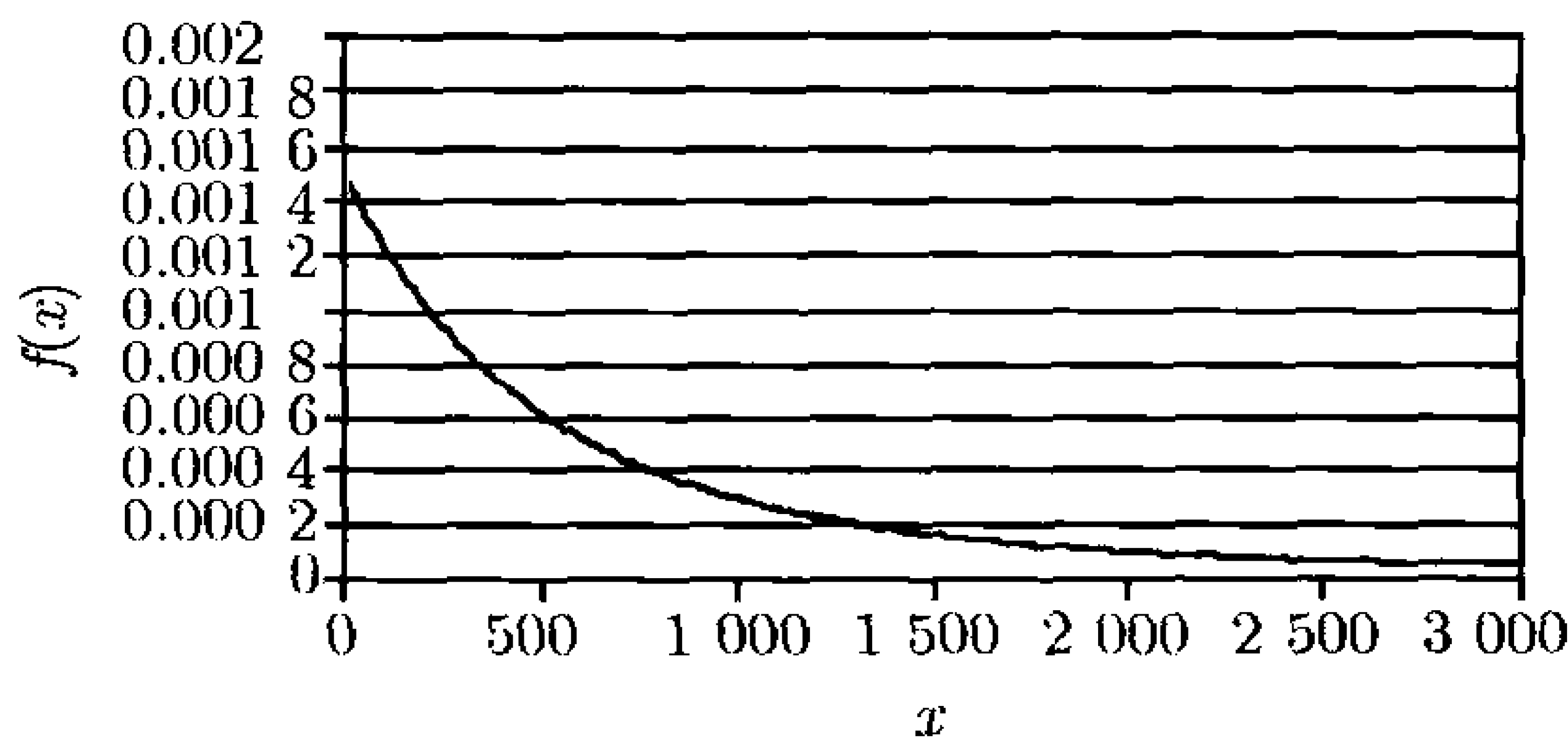


图 2-6 模型 2 的密度函数

定义 2.12 随机变量X的概率函数, 也称为概率质点函数通常记为$p_X(x)$或$p(x)$, 表示随机变量X在概率值非零点的概率, 正式的定义为: $p_X(x)=\Pr(X=x)$.

对于离散型随机变量, 自然有如下分布函数和生存函数关于概率函数的表达式: $F(x)=\sum\limits_{y\leqslant x} p(y)$ 和 $S(x)=\sum\limits_{y>x} p(y)$.

例 2.13 给出前面 4 个模型的概率函数.

解　对 4 个模型来说,

$$p_1(x)\text{未定义},$$
$$p_2(x)\text{未定义},$$
$$p_3(x)=\begin{cases}0.50, & x=0,\\ 0.25, & x=1,\\ 0.12, & x=2,\\ 0.08, & x=3,\\ 0.05, & x=4,\end{cases}$$
$$p_4(0)=0.7.$$

再次注意模型 4 为混合型, 所以上面的概率函数只描述了离散部分. 很难用一种简单的方法统一给出混合分布的概率或密度函数. 另一方面, 对于混合模型应该主要刻画其混合的特征, 对于模型 4 我们可以如下表示概率密度函数:

$$f_4(x)=\begin{cases}0.7, & x=0,\\ 0.000\,003\mathrm{e}^{-0.000\,01x}, & x>0.\end{cases}$$

实际上, 从技术的角度看这个函数根本不是概率密度函数. 当我们对概率密度函数在某个点赋予正实数值而不是定义在某个区间上时, 可以认为这个值是一个离散的概率质点. □

定义 2.14　*随机变量X的风险率, 也称为死亡力或者失效率, 通常记为$h_X(x)$或$h(x)$, 表示在随机变量X的密度函数有定义的点, 其密度函数与生存函数的比值, 即 $h_X(x)=f_X(x)/S_X(x)$.*

在被称为死亡力时, 风险率常常表示为 $\mu(x)$; 在被称为失效率时, 风险率常常表示为 $\lambda(x)$. 无论怎样, 它都是表示已知某个随机变量至少已达到 x 时它在点 x 发生的概率密度. 又由 $h_X(x)=-S'(x)/S(x)=-\mathrm{d}\ln S(x)/\mathrm{d}x$, 生存函数可以表示为 $S(b)=\mathrm{e}^{-\int_0^b h(x)\mathrm{d}x}$. 虽然不一定是必需的, 但是从这个表达式看出其支集为非负实数. 按照生存分析的术语, 死亡力表示已知年龄为 x 的个体在今后一年内死亡的概率, 因此是一种年度化的概率[①], 也就是每年的死亡率. 在本书中, 我们统一用 $h(x)$ 表示风险率, 尽管可能会有不同的名称.

例 2.15　*给出前面 4 个模型的风险率函数.*

解　对 4 个模型来说,

① 注意, 死亡力不是概率, 特别地, 它的值可能会大于 1, 当然把它看作是一个概率也无妨.

$$h_1(x) = \frac{0.01}{1-0.01x}, \quad 0 < x < 100,$$

$$h_2(x) = \frac{3}{x+2\ 000}, \quad x > 0,$$

$h_3(x)$未定义,

$$h_4(x) = 0.000\ 01, \qquad x > 0.$$

同样地, 模型 4 的混合型随机变量的风险率函数只是在随机变量支集的一部分上有定义, 这一点与前面的情况不同, 这里既涉及概率密度函数也涉及概率函数. 在具有离散概率质点的部分, 风险率函数没有定义. □

例 2.16 给出模型 1 和模型 2 的风险率函数的图形.

解 见图 2-7 和图 2-8. □

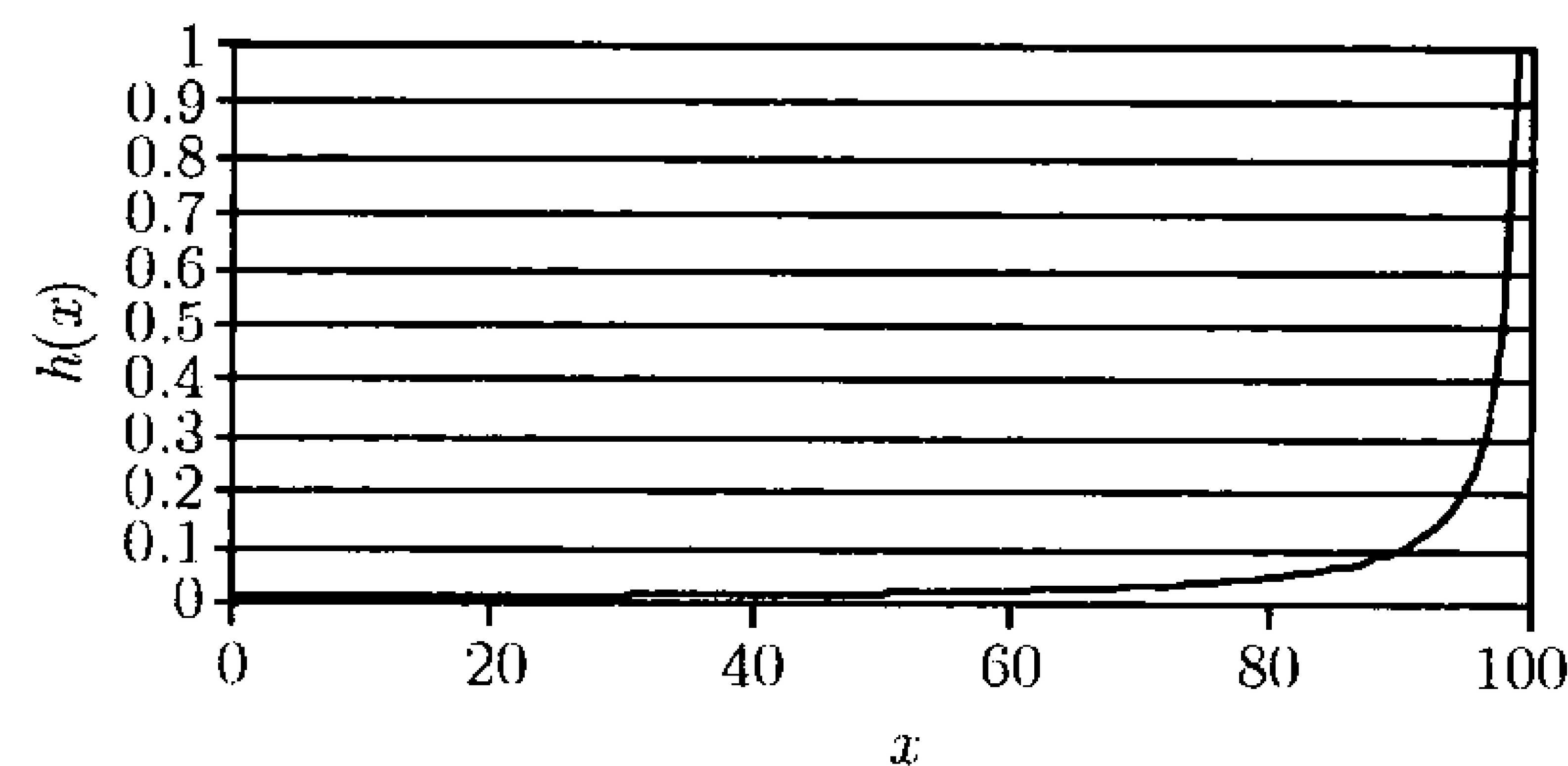


图 2-7 模型 1 的风险率函数

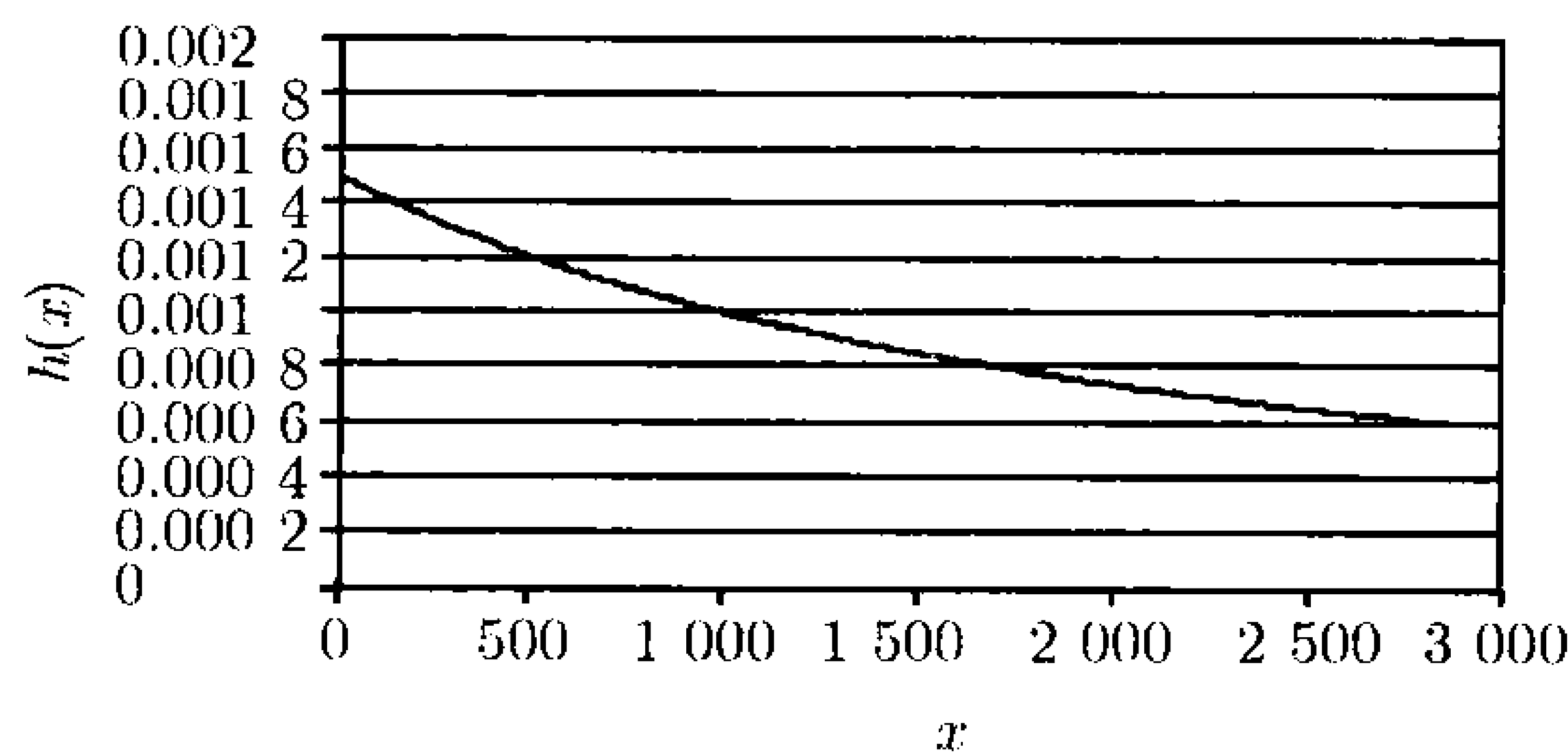


图 2-8 模型 2 的风险率函数

下面介绍的模型考虑了密度函数和风险率函数都没有定义的特殊情况.

模型 5 这里将模型 1 描述的简单寿命分布替换为另一种形式. 这个函数是分段线性的, 而且

$$S_5(x) = \begin{cases} 1-0.01x, & 0 \leqslant x < 50, \\ 1.5-0.02x, & 50 \leqslant x < 75. \end{cases}$$

在 50 这个点没有导数. 因此, 无论是密度函数还是风险率函数在 50 这个点都没有定义. 与模型 4 代表的混合型随机变量不同, 这个模型在 50 这个点没有离散的概

率质点. 由于随机变量取 50 的概率为零, 因此, 我们可以在 50 这个点任意给密度函数或风险率函数一个值, 这不会影响其他的计算结果. 这个值可以任意定义, 以使得这个函数是右连续①的. 习题 2.1 的解可为此例.

本书的附录 A 给出了常用的连续型分布函数, 附录 B 给出了常用的离散型分布函数. 随机变量的一个重要特征是其最有可能发生的值.

定义 2.17 随机变量X的众数是指其最有可能发生的值, 对于离散型随机变量这个点是概率函数的最大值点, 对于连续型随机变量这个点是密度函数的最大值点. 如果有局部的最大值点, 这些点也是众数.

例 2.18 给出前面 5 个模型的众数.

解 对于模型 1, 密度函数为常数, 可以说从 0 到 100 的所有点都是众数, 也可以说, 这个随机变量没有众数.

对于模型 2 密度函数严格递减, 所以众数为 0 点.

对于模型 3 在 0 点的概率最大.

模型 4 作为一种混合随机变量, 无法定义众数.

对于模型 5 在两个区间内的密度函数为常数, 其值分别为 50 和 75, 所以, 这两个值都是众数. □

习题

2.1 给出模型 5 的分布函数、密度函数和风险率函数.

2.2 给出模型 3~5 的分布函数图形. 同时对于存在下述函数的模型绘出图形: 密度函数或概率函数、风险率函数.

2.3* 已知随机变量 X 的密度函数为 $f(x) = 4x(1+x^2)^{-3}$, $x > 0$. 试确定 X 的众数.

2.4* 某非负随机变量的风险率函数为 $h(x) = A + e^{2x}$, $x \geqslant 0$. 而且已知 $S(0.4) = 0.5$, 试确定 A 的值.

2.5* 已知随机变量 X 服从参数为 $\alpha = 2$ 和 $\theta = 10\ 000$ 的 Pareto 分布, 随机变量 Y 服从参数为 $\alpha = 2, \gamma = 2$ 和 $\theta = \sqrt{20\ 000}$ 的 Burr 分布. 记 r 为 $\Pr(X > d)$ 与 $\Pr(Y > d)$ 的比值, 计算 $\lim\limits_{d\to\infty} r$.

① 通过在这个点上任意定义密度函数或风险率函数的值, 均可计算生存函数. 如果在这个点上有离散概率 (在这种情境这些点的值不能任意定义), 则密度函数或风险率函数都不足以刻画分布函数.

第 3 章　分布函数的数字特征

3.1　矩

我们可以对前面介绍的模型进行各种有意义的计算. 例如, 对某个给定的免赔额或保单限额, 计算每次索赔的平均赔付, 也可以计算当前年龄为 40 岁的个体的平均未来生存时间.

定义 3.1　*定义随机变量的k阶原点矩(raw moment)为该随机变量k次幂的期望 (平均) 值, 前提条件是这个期望值存在. 用$\mathrm{E}[X^k]$ 或 μ_k'表示. 称一阶原点矩为随机变量的均值(mean), 通常记为μ.*

注意, 这里的 μ 与前面的死亡力 $\mu(x)$ 不是一个概念. 对于非负随机变量 (即 $\Pr(X \geqslant 0) = 1$), k 可以是任何实数. 在推导这些量的计算公式时, 有必要区分连续与离散变量. 下面将给出处处连续或者处处离散的随机变量的计算公式. 对于混合模型计算数字特征, 首先需要分别对连续部分的密度函数进行积分计算和对离散部分的概率函数进行求和计算, 然后将结果相加, 得到总的数值. k 阶原点矩的计算公式如下:

$$\mu_k' = \mathrm{E}(X^k) = \begin{cases} \int_{-\infty}^{\infty} x^k f(x)\mathrm{d}x, & \text{若随机变量为连续的,} \\ \sum\limits_j x_j^k p(x_j), & \text{若随机变量为离散的,} \end{cases} \tag{3.1}$$

这里只对所有具有严格正概率的 x_j 求和. 最后, 应注意积分式或求和式有可能不收敛, 在这种情况下称 k 阶原点矩不存在.

例 3.2　*试计算下面 5 个模型的前两阶原点矩.*

解　用随机变量 X 的下标表示它所代表的模型.

$$\begin{aligned}
\mathrm{E}(X_1) &= \int_0^{100} x(0.01)\mathrm{d}x = 50, \\
\mathrm{E}(X_1^2) &= \int_0^{100} x^2(0.01)\mathrm{d}x = 3\ 333.33, \\
\mathrm{E}(X_2) &= \int_0^{\infty} x\frac{3(2\ 000)^3}{(x+2\ 000)^4}\mathrm{d}x = 1\ 000, \\
\mathrm{E}(X_2^2) &= \int_0^{\infty} x^2\frac{3(2\ 000)^3}{(x+2\ 000)^4}\mathrm{d}x = 4\ 000\ 000, \\
\mathrm{E}(X_3) &= 0(0.5) + 1(0.25) + 2(0.12) + 3(0.08) + 4(0.05) = 0.93,
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}(X_3^2) &= 0(0.5)+1(0.25)+4(0.12)+9(0.08)+16(0.05)=2.25,\\
\mathrm{E}(X_4) &= 0(0.7)+\int_0^{\infty} x(0.000\,003)\mathrm{e}^{-0.000\,01x}\mathrm{d}x = 30\,000,\\
\mathrm{E}(X_4^2) &= 0^2(0.7)+\int_0^{\infty} x^2(0.000\,003)\mathrm{e}^{-0.000\,01x}\mathrm{d}x = 6\,000\,000\,000,\\
\mathrm{E}(X_5) &= \int_0^{50} x(0.01)\mathrm{d}x+\int_{50}^{75} x(0.02)\mathrm{d}x = 43.75,\\
\mathrm{E}(X_5^2) &= \int_0^{50} x^2(0.01)\mathrm{d}x+\int_{50}^{75} x^2(0.02)\mathrm{d}x = 2\,395.83.
\end{aligned}$$ □

在进一步讨论之前, 先介绍一个新的模型. 它类似于模型 3, 但是有一些本质的区别, 这是一个离散的模型, 而且要求所有的概率值都是某个数的整数倍. 另外, 这个模型将以某种特殊的方式与样本数据建立联系.

定义 3.3 经验分布模型(empirical)是基于一个样本量为 n 并对每个数据点赋予概率 $1/n$ 的离散分布模型.

模型 6 考虑有 8 个样本的观测数据: 3, 5, 6, 6, 6, 7, 7, 10. 其经验分布模型的概率函数为:

$$p_6(x)=\begin{cases}0.125, & x=3,\\ 0.125, & x=5,\\ 0.375, & x=6,\\ 0.25, & x=7,\\ 0.125, & x=10.\end{cases}$$ □

细心的读者将发现很多取值有限的离散模型看上去很像是经验分布模型. 模型 3 就可以被看作是 100 个样本的经验模型, 其数据为: 50 个 0, 25 个 1, 12 个 2, 8 个 3, 5 个 4. 无论如何, 我们只是对基于真实样本数据时的模型采用经验分布模型这个术语. 采用模型 3 中使用的方法, 模型 6 的前两阶原点矩为

$$\mathrm{E}(X_6)=6.25,\quad \mathrm{E}(X_6^2)=42.5.$$

值得注意的是, 这个模型的随机变量的均值等于样本的算术平均值 (也称为样本均值).

定义 3.4 定义随机变量的k阶中心矩(central moment)为该变量与其均值的偏差的k次幂的期望值, 一般表示为$\mathrm{E}[(X-\mu)^k]$或μ_k. 通常称二阶中心矩为方差(variance), 一般用σ^2表示, 它的平方根σ称为标准差(standard deviation). 标准差与均值的比值称为变异系数(coefficient variance). 三阶中心矩与标准差的立方的比值$\gamma_1=\mu_3/\sigma^3$, 称为偏度(skewness). 四阶中心矩与标准差的四次方的比值$\gamma_2=\mu_4/\sigma^4$被称为峰度(kurtosis)[①].

① 更为精确地说, 这些特征值应该为“偏度系数”和“峰度系数”, 因为还有其他的量也可以度量对称性和平坦性. 但本书将始终采用这种简单的表达式.

对于完全连续或完全离散情形, 中心矩的计算公式为

$$\mu_k = \mathrm{E}\left[(X-\mu)^k\right] = \begin{cases} \int_{-\infty}^{\infty}(x-\mu)^k f(x)\mathrm{d}x, & \text{若随机变量为连续的,} \\ \sum\limits_j (x_j-\mu)^k p(x_j), & \text{若随机变量为离散的.} \end{cases} \tag{3.2}$$

事实上, 仅需要对 $f(x)$ 为正的 x 进行积分. 标准差是对随机变量在其可能值上概率分散程度的一个度量, 它与随机变量的度量单位一致. 变异系数衡量了上述标准差相对于均值的分散程度. 偏度是关于对称性的度量, 一个完全对称的分布其偏度为 0, 偏度为正表示: 概率相同时, 与左边的值相比, 右边的值将离均值更远. 峰度度量了相对于正态分布的平坦程度 (正态分布的峰度为 3). 当标准差相同时, 峰度大于 3 的分布相对于正态分布在远离均值的点的概率更大. 变异系数、偏度和峰度都是无量纲的.

原点矩与中心矩之间存在一定的联系, 下面的等式揭示了两者之间的关系. 推导过程中使用了 (3.1) 和 (3.2) 的连续情形, 但结果适用于所有类型的随机变量.

$$\begin{aligned} \mu_2 &= \int_{-\infty}^{\infty}(x-\mu)^2 f(x)\mathrm{d}x = \int_{-\infty}^{\infty}(x^2-2x\mu+\mu^2)f(x)\mathrm{d}x \\ &= \mathrm{E}(X^2)-2\mu\mathrm{E}(X)+\mu^2 = \mu_2'-\mu^2. \end{aligned} \tag{3.3}$$

例 3.5 试证明gamma(伽玛) 分布的密度函数是正偏的, 并作图举例说明.

解 由附录 A 可知 gamma 分布的前三阶原点矩分别为: $\alpha\theta$, $\alpha(\alpha+1)\theta^2$, $\alpha(\alpha+1)(\alpha+2)\theta^3$. 由 (3.3) 式得到 gamma 分布的方差为 $\alpha\theta^2$, 由习题 3.1 可知三阶中心矩为 $2\alpha\theta^3$. 因此, 偏度为 $2\alpha^{-1/2}$. 因为 α 必须为正数, 所以偏度总为正. 同样, 随着 α 的减少, 偏度会增大.

考虑如下两个具体的 gamma 分布. 一个的参数为 $\alpha=0.5$, $\theta=100$; 另一个的参数为 $\alpha=5$, $\theta=10$. 两个分布有相同的均值, 但偏度系数分别为 2.83 和 0.89. 图 3-1 显示了两个分布的不同之处. □

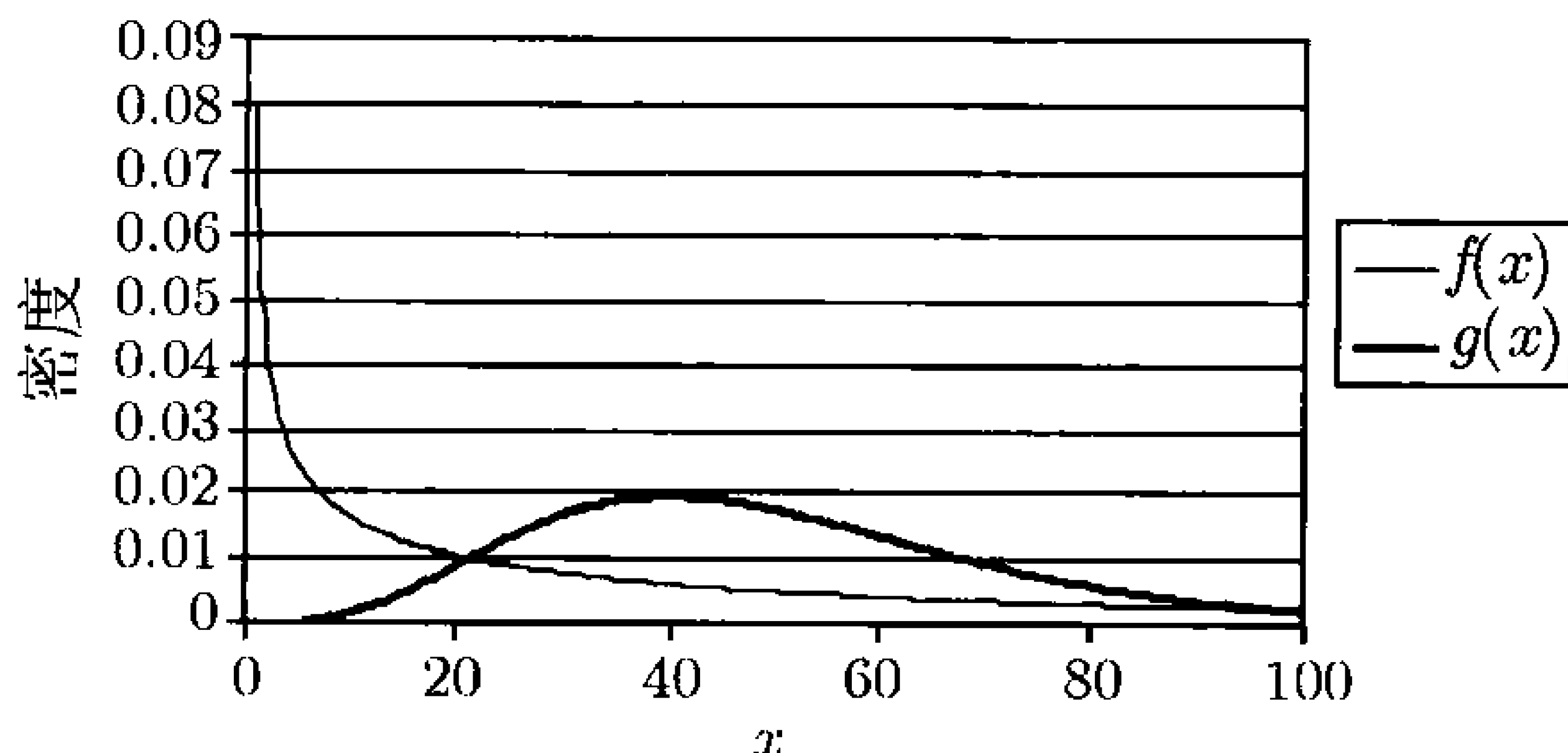


图 3-1 $f(x)\sim \text{gamma}(0.5,100)$ 和 $g(x)\sim \text{gamma}(5,10)$ 的密度函数

注意, 在习题 3.2 计算模型 6 的标准差时, 分母为 n(更常用的是 $n-1$). 最后, 还应注意, 在计算各种原点矩时, 积分或者求和可能并不存在 (例如模型 2 的三阶矩和四阶矩). 由于我们常见模型的被积函数或者求和数都是非负的, 所以矩不存在就意味着积分或者求和式的极限为正无穷. 具体见例 4.15 的说明.

定义 3.6 对于满足$\Pr(X>d)>0$的给定值d, 定义超损变量(excess loss variable)为$Y=X-d$若$X>d$. 该变量的期望

$$e_X(d)=e(d)=\mathrm{E}(Y)=\mathrm{E}(X-d|X>d),$$

称为平均超损函数(mean excess loss variable). 也称这个期望值为平均未来寿命函数(mean residual life function)和完全未来寿命的期望(complete expectation of life). 对后者通常使用记号$\overset{0}{e}_d$.

也称这个变量为**左截断平移变量**(left truncated and shifted variable). 称之左截断是因为 d 以下的观测都被丢弃了, 平移是指所有的观测量都减去了 d. 若 X 表示赔付变量, 则平均超损为当赔付额超过免赔额 d 时赔付额的期望. 若 X 表示死亡年龄, 则平均超损考虑了那些在 d 岁时仍存活的个体未来的剩余寿命的期望. 下式给出了超损变量的 k 阶矩:

$$e_X^k(d)=\begin{cases}\dfrac{\int_d^\infty (x-d)^k f(x)\mathrm{d}x}{1-F(d)}, & \text{若随机变量为连续的,}\\[2ex] \dfrac{\sum_{x_j>d}(x_j-d)^k p(x_j)}{1-F(d)}, & \text{若随机变量为离散的.}\end{cases} \tag{3.4}$$

这里, 仅当上述积分或者求和式收敛的时候 $e_X^k(d)$ 才有定义. 一阶矩的计算有特别简便的公式. 下面对连续情形进行推导, 但其结果对所有随机变量都成立. 其中的第二行由分部积分得到, 并记 $f(x)$ 的不定积分为 $-S(x)$.

$$\begin{aligned} e_X(d)&=\frac{\int_d^\infty (x-d)f(x)\mathrm{d}x}{1-F(d)}\\ &=\frac{-(x-d)S(x)|_d^\infty+\int_d^\infty S(x)\mathrm{d}x}{S(d)}\\ &=\frac{\int_d^\infty S(x)\mathrm{d}x}{S(d)}. \end{aligned} \tag{3.5}$$

定义 3.7 定义左删失平移变量为

$$Y=(X-d)_+=\begin{cases}0, & X<d,\\ X-d, & X\geqslant d.\end{cases}$$

因为没有将小于 d 的值忽略而是定义为 0, 所以称之为左删失. 这个变量的各种矩并没有标准的术语或者符号. 对于以货币计量的损失事件, 超损变量和左删失

平移变量的区别在于前者表示**平均赔付量**, 后者表示**平均损失量**. 只有赔付发生时, 前一种情况的变量才会存在; 而在后一种情况, 若损失不造成赔付, 则变量取值为 0. 左删失平移变量各阶矩的计算公式如下

$$\mathrm{E}[(X-d)_+^k]=\begin{cases}\int_d^{\infty}(x-d)^k f(x), & \text{若随机变量是连续的,}\\ \sum\limits_{x_j>d}(x_j-d)^k p(x_j), & \text{若随机变量是离散的.}\end{cases} \tag{3.6}$$

还应该注意到

$$\mathrm{E}[(X-d)_+^k]=e^k(d)[1-F(d)]. \tag{3.7}$$

例 3.8 试作图说明超损变量和左删失平移变量之间的不同.

解 在图 3-2 和图 3-3 中绘制了修正的随机变量 Y 作为未修正的随机变量 X 的函数关系. 两个图的唯一不同是 X 小于 100 时 Y 的定义: 超损变量此时没有定义, 而左删失平移变量在这种情况下定义为 0. □

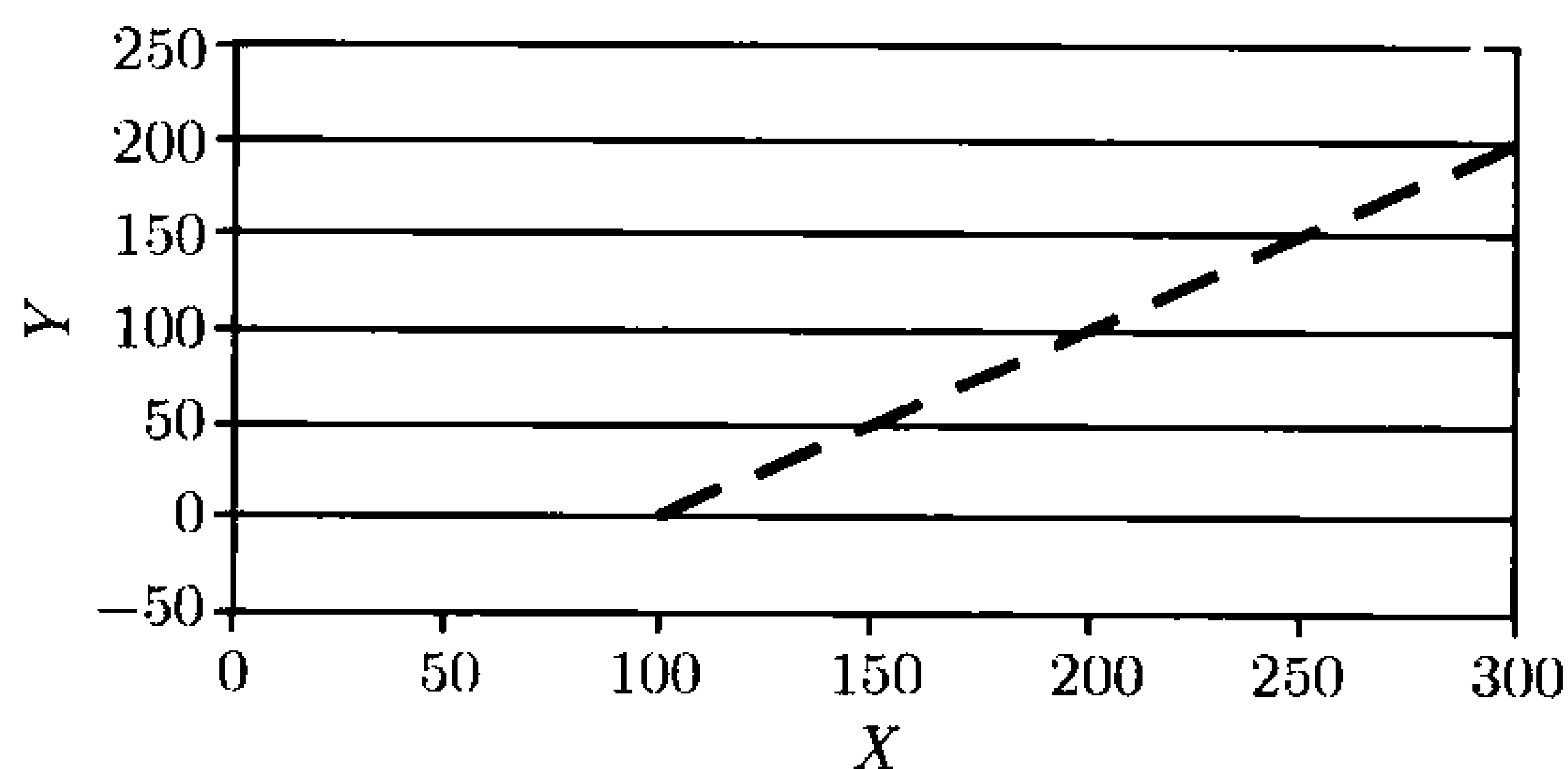


图 3-2 超损变量

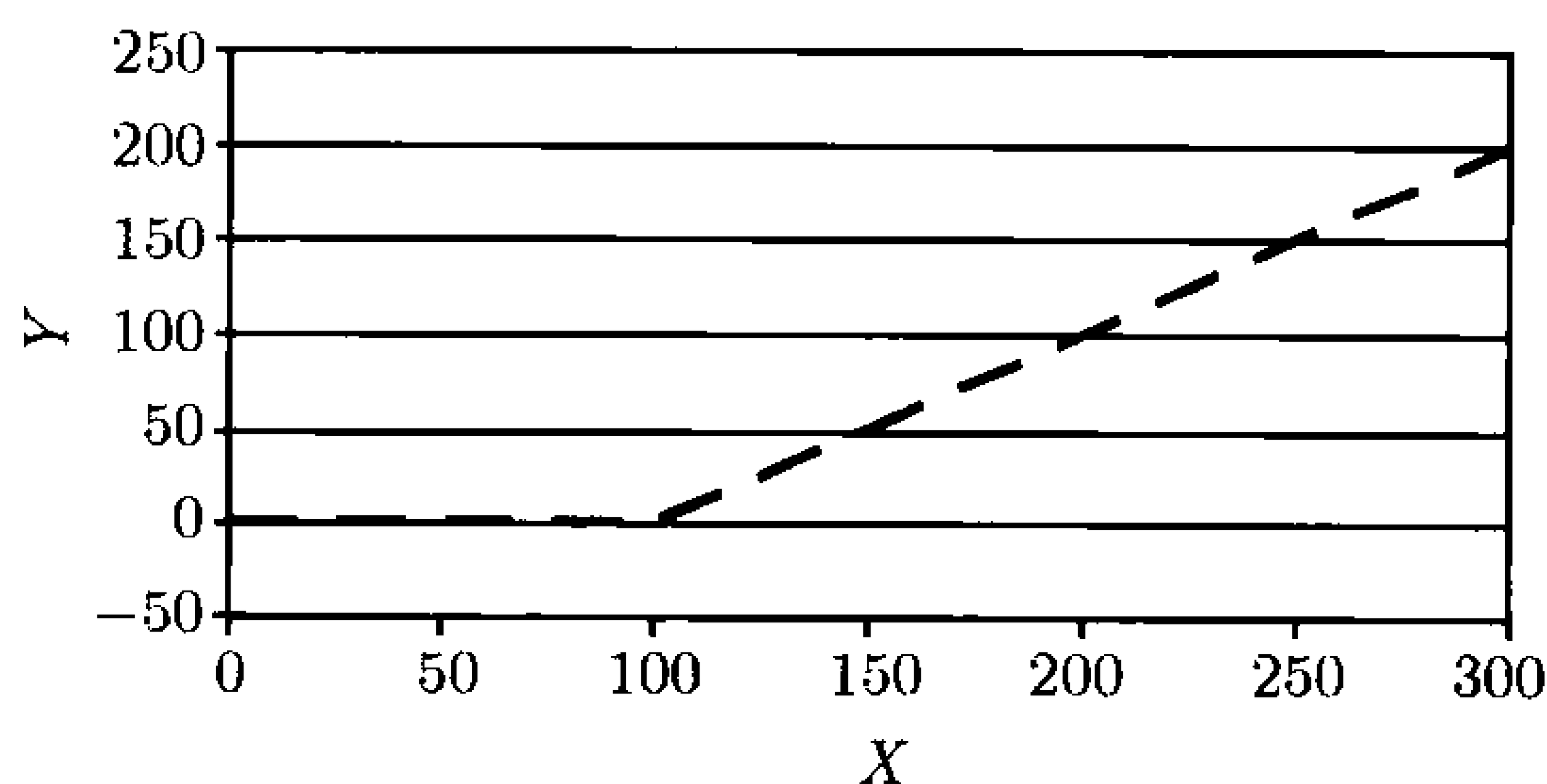


图 3-3 左删失平移变量

下面的定义给出了超损函数的余函数.

定义 3.9 定义限额损失变量为

$$Y=X\wedge u=\begin{cases}X, & X<u,\\ u, & X\geqslant u.\end{cases}$$

它的数学期望值 $\mathrm{E}[X \wedge u]$ 称为**限额期望值**.

因为大于 u 的部分都取为 u, 所以也可以称这个变量为**右删失变量**. 与该变量有关的一个保险现象是许多的保险产品都规定了赔付的最大限额. 实际上, 有 $(X-d)_+ + (X \wedge d) = X$. 即在购买一份最大赔付额为 d 的保单的同时购买一份免赔额为 d 的保单相当于购买了全额保险. 如图 3-4 所示.

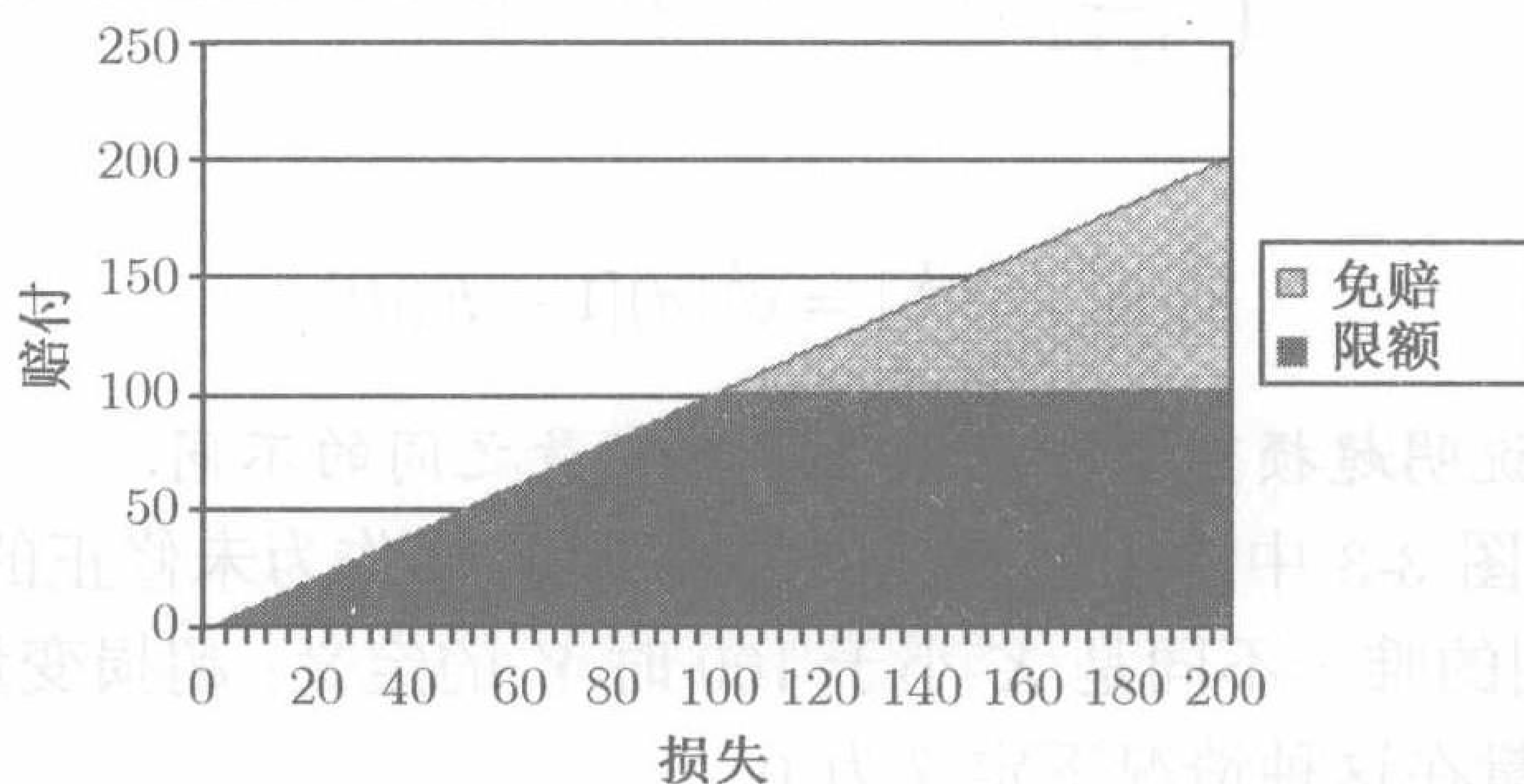


图 3-4　限额 100 元与免赔 100 元的组合等于全额赔付

限额变量 k 阶矩的最直接计算公式为

$$\mathrm{E}[(X \wedge u)^k] = \begin{cases} \int_{-\infty}^{u} x^k f(x)\mathrm{d}x + u^k[1-F(u)], & \text{若随机变量是连续的,} \\ \sum\limits_{x_j \leqslant u} x_j^k p(x_j) + u^k[1-F(u)], & \text{若随机变量是离散的.} \end{cases} \tag{3.8}$$

另一个有意义的公式由下面的推导得到

$$\begin{aligned} \mathrm{E}[(X \wedge u)^k] &= \int_{-\infty}^{0} x^k f(x)\mathrm{d}x + \int_0^u x^k f(x)\mathrm{d}x + u^k[1-F(u)] \\ &= x^k F(x)_{-\infty}^{0} - \int_{-\infty}^{0} kx^{k-1}F(x)\mathrm{d}x \\ &\quad - x^k S(x)_0^u + \int_0^u kx^{k-1}S(x)\mathrm{d}x + u^k S(u) \\ &= -\int_{-\infty}^{0} kx^{k-1}F(x)\mathrm{d}x + \int_0^u kx^{k-1}S(x)\mathrm{d}x, \end{aligned} \tag{3.9}$$

其中的第二行由分部积分得到. 对 $k=1$, 有

$$\mathrm{E}(X \wedge u) = -\int_{-\infty}^{0} F(x)\mathrm{d}x + \int_0^u S(x)\mathrm{d}x.$$

相应的离散型随机变量的类似公式就没有那种特别的意义了. 限额损失的期望值也代表了当免赔被强制执行时每件赔案所节省的支付的期望值. 附录 A 中列出了一些常用的连续分布的限额变量的 k 阶矩. 习题 3.8 要求推导前面介绍的 3 个一阶矩.

习题

3.1 类似于公式 (3.3), 试推导 μ_3 和 μ_4 的公式.

3.2 分别计算 6 个模型的标准差、偏度和峰度. 注意: 模型 2 是 Pareto 分布, 模型 4 密度函数的连续部分为指数分布. 附录 A 将对这些计算有一定的帮助.

3.3* 已知某随机变量的均值和变异系数均为 2, 三阶原点矩为 136, 试求偏度.

3.4* 求变异系数为 1 的 gamma 分布的偏度.

3.5 试计算模型 1 到模型 4 的平均超损函数, 并对模型 1、2 和 4 的函数表达式进行比较.

3.6* 现有两个随机变量 X 和 Y, $e_Y(30) = e_X(30) + 4$. X 服从区间 $[0, 100]$ 上的均匀分布, Y 服从 $[0, w]$ 上的均匀分布, 求 w.

3.7* 已知某随机变量的密度函数 $f(x) = \lambda^{-1}\mathrm{e}^{-x/\lambda}$, $x, \lambda > 0$. 试计算 λ 岁的平均未来寿命函数 $e(\lambda)$.

3.8 证明下式成立:

$$\mathrm{E}(X) = e(d)S(d) + \mathrm{E}(X \wedge d). \tag{3.10}$$

3.9 利用公式 (3.8) 和 (3.10) 求模型 1~4 的限额期望值函数, 对模型 1 和模型 2 也可以利用公式 (3.9).

3.10* 下面陈述中正确的为

(a) 经验分布的平均未来寿命函数是连续的.

(b) 指数分布的平均未来寿命函数为常数.

(c) 如果 Pareto 分布的平均未来寿命函数存在, 则为递减函数.

3.11* 已知某损失服从 $\alpha = 0.5$ 和 $\theta = 10\,000$ 的 Pareto 分布. 求 10 000 点的平均未来寿命.

3.12 定义一个右截断变量并给出它的 k 阶矩公式.

3.13* 已知某个体赔案分布的概率密度函数为

$$f(x) = 2.5x^{-3.5}, \quad x \geqslant 1.$$

求变异系数.

3.14* 已知如下的赔案记录: 100, 200, 300, 400, 500, 这些值的真实概率分别为: 0.05, 0.20, 0.50, 0.20, 0.05. 求这个分布的偏度和峰度.

3.15* 已知某损失服从参数 $\alpha > 1$ 和 θ 未知的 Pareto 分布, 求 $x = 2\theta$ 点的平均超损函数与 $x = \theta$ 点的平均超损函数的比值.

3.2 分 位 数

能够由分布函数得到的另一个有意义的特征值为分位数函数, 它是分布函数的逆函数, 但是这个量并不是总能够给出其唯一定义, 所以必须建立一种任意情况都可行的定义.

定义 3.10 称任何满足 $F(\pi_p-) \leqslant p \leqslant F(\pi_p)$的$\pi_p$为随机变量的 p 分位数. 又称 50% 分位数 $\pi_{0.5}$为随机变量的中位数.

如果某分布函数有且仅有一个 x 的函数值为 p, 则分位数是唯一确定的. 若分布函数从小于 p 的值跳到大于 p 的值, 则分位数就是这个跳跃点的值. 分位数不唯一的情形只能是分布函数在某个范围内 (不止一个点) 的取值都是 p, 在这种情形下, 这个范围内的任何 x 的取值都可以被认为是 p 分位数.

例 3.11 求模型 1 和模型 3 的 50% 分位数和 80% 分位数.

解 模型 1 的 p 分位数可以通过求解 $p = F(\pi_p) = 0.01\pi_p$ 得到, 所以 $\pi_p = 100p$. 所求的 50% 和 80% 分位数见图 3-5.

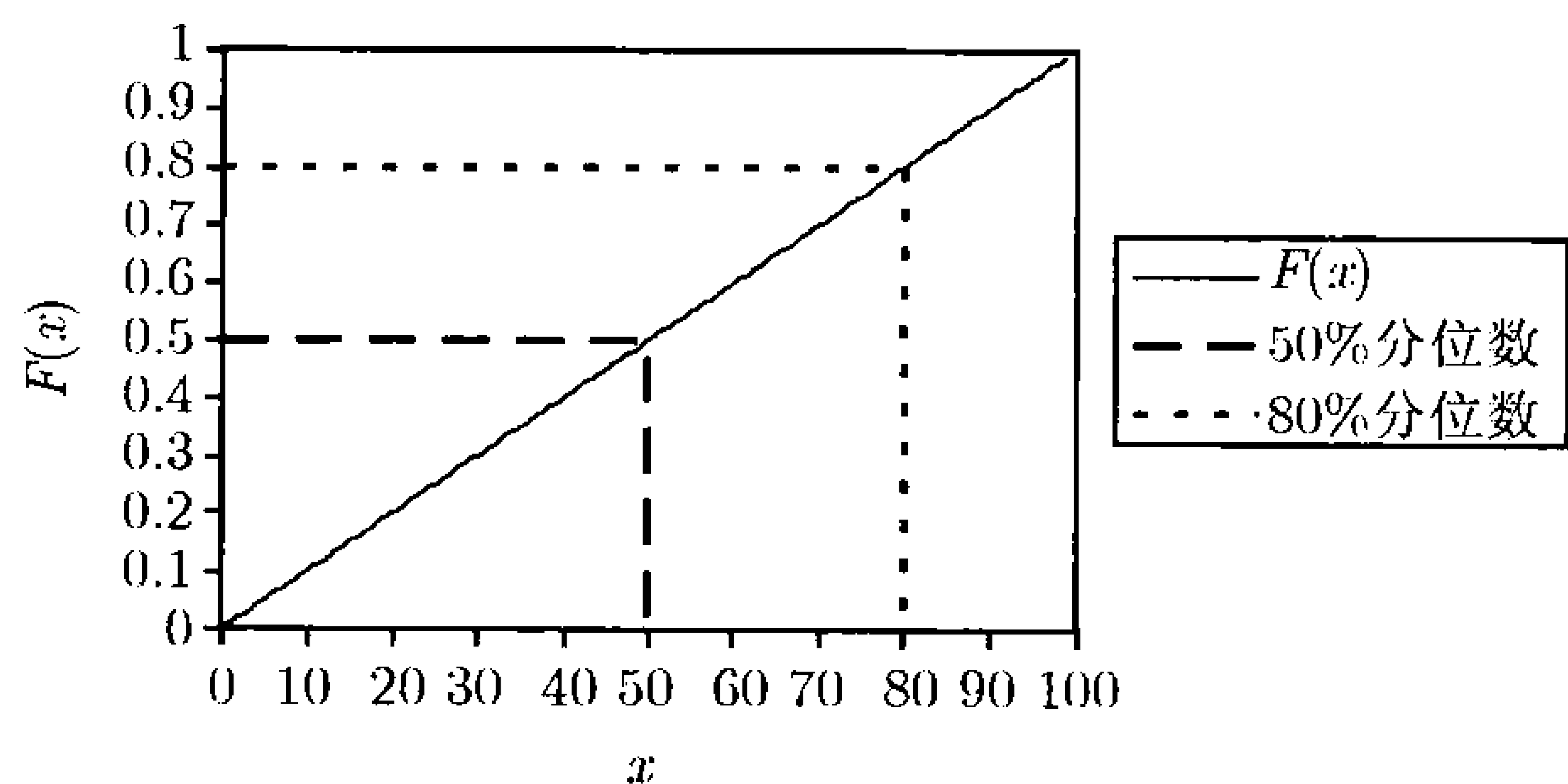


图 3-5 模型 1 的分位数

模型 3 的分布函数对所有的 $0 \leqslant x < 1$ 都等于 0.5, 所有这些值都可以作为 50% 分位数. 对于 80% 分位数, 在 $x = 2$ 这个点, 分布函数从 0.75 直接跳到 0.87, 所以 $\pi_{0.8} = 2$(见图 3-6).

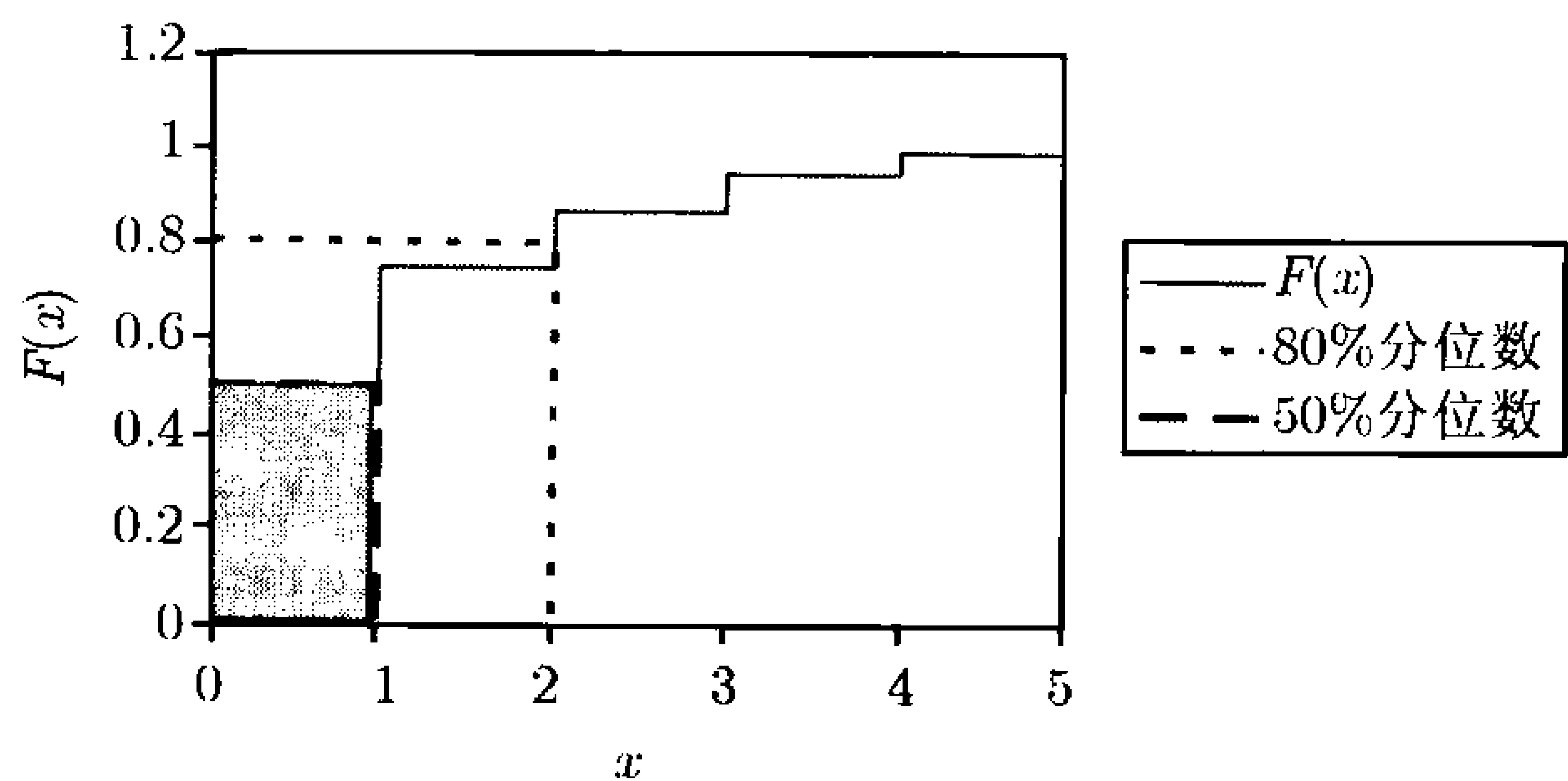


图 3-6 模型 3 的分位数

□

习题

3.16* 已知某随机变量的累积分布函数为 $F(x) = 1 - x^{-2}, x \geqslant 1$. 求随机变量的均值、中位数和众数.

3.17 求模型 2 与模型 4~6 的 50%分位数和 80%分位数.

3.18* 已知某损失服从参数为 α 和 θ 的 Pareto 分布, 10%分位数是 $\theta - k$, 90%分位数是 $5\theta - 3k$. 求 α 的值.

3.19* 已知某损失服从参数为 τ 和 θ 的 Weibull(威布尔) 分布, 25%分位数是 1 000, 75%分位数是 100 000. 求 τ 的值.

3.3 生成函数与随机变量和

现实中保险公司很少只承保一个个体, 所有保单的总赔付是所有支付的总和. 因此, 能够确定 $S_k = X_1 + \cdots + X_k$ 的性质将是非常有用的. 下面的第一个结论是中心极限定理的一种形式.

定理 3.12 对如上定义的随机变量 S_k, 有 $\mathrm{E}(S_k) = \mathrm{E}(X_1) + \cdots + \mathrm{E}(X_k)$. 同样地, 若$X_1, \cdots, X_k$独立, 则有$\mathrm{Var}(S_k) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_k)$. 若随机变量$X_1, X_2 \cdots$独立且前两阶矩符合一定的条件, 则$\lim\limits_{k\to\infty} [S_k - \mathrm{E}(S_k)]/\sqrt{\mathrm{Var}(S_k)}$服从均值为 0, 方差为 1 的正态分布.

要得到 S_k 的分布函数或密度函数通常很难, 但是, 也有一些极少的情形会有比较简单的结果. 要得到这种简化关键是生成函数.

定义 3.13 对随机变量X, 定义矩母函数(moment generating function, mgf)为如下关于t的函数: $M_X(t) = \mathrm{E}(\mathrm{e}^{tX})$, 只当期望存在时有定义. 定义概率生成函数(probability generating function, pgf)为如下关于z的函数: $P_X(z) = \mathrm{E}(z^X)$, 只当期望存在时有定义.

注意 $M_X(t) = P_X(\mathrm{e}^t)$, $P_X(z) = M_X(\ln z)$. 通常 mgf 用于连续随机变量, pgf 用于离散随机变量. 对我们来说, 这些函数的价值不在于它们是随机变量矩的函数或者概率的函数, 而是随机变量的分布函数与它们的 mgf 和 pgf 之间的一一对应关系 (例如, 两个分布函数不同的随机变量不可能有相同的 mgf 或 pgf). 下面的结论将有助于计算随机变量和的分布.

定理 3.14 令$S_k = X_1 + \cdots + X_k$, 求和式中各项随机变量独立, 而且各随机变量的 mgf 或者 pgf 存在, 则有

$$M_{S_k}(t) = \prod_{j=1}^{k} M_{X_j}(t), \quad P_{S_k}(z) = \prod_{j=1}^{k} P_{X_j}(z).$$

证明 利用独立随机变量乘积的期望为各个随机变量期望的乘积这一结论, 则有

$$\begin{aligned} M_{S_k}(t) &= \mathrm{E}(\mathrm{e}^{tS_k}) = \mathrm{E}[\mathrm{e}^{t(X_1+\cdots+X_k)}] \\ &= \prod_{j=1}^{k} \mathrm{E}(\mathrm{e}^{tX_j}) = \prod_{j=1}^{k} M_{X_j(t)}. \end{aligned}$$

类似的证明也适用于 pgf. □

例 3.15　证明独立 gamma 分布随机变量(θ参数值相同)之和仍然为 gamma 分布.

解　gamma 分布的矩母函数是

$$\begin{aligned}
\mathrm{E}(\mathrm{e}^{tX}) &= \frac{\int_0^{\infty} \mathrm{e}^{tx} x^{\alpha-1} \mathrm{e}^{-x/\theta} \mathrm{d}x}{\Gamma(\alpha)\theta^{\alpha}} \\
&= \frac{\int_0^{\infty} x^{\alpha-1} \mathrm{e}^{-x(-t+1/\theta)} \mathrm{d}x}{\Gamma(\alpha)\theta^{\alpha}} \\
&= \frac{\int_0^{\infty} y^{\alpha-1} (-t+1/\theta)^{-\alpha} \mathrm{e}^{-y} \mathrm{d}y}{\Gamma(\alpha)\theta^{\alpha}} \\
&= \frac{\Gamma(\alpha)(-t+1/\theta)^{-\alpha}}{\Gamma(\alpha)\theta^{\alpha}} = \left(\frac{1}{1-\theta t}\right)^{\alpha}, \quad t < 1/\theta.
\end{aligned}$$

令 X_j 服从参数为 α_j 和 θ 的 gamma 分布, 则求和后的矩母函数为

$$M_{S_k}(t) = \prod_{j=1}^{k} \left(\frac{1}{1-\theta t}\right)^{\alpha_j} = \left(\frac{1}{1-\theta t}\right)^{\alpha_1+\cdots+\alpha_k}$$

这个表达式是参数为 $\alpha_1 + \cdots + \alpha_k$ 和 θ 的 gamma 分布的矩母函数. □

例 3.16　求 Poisson(泊松) 分布的 mgf 和 pgf.

解　pgf 为

$$P_X(z) = \sum_{x=0}^{\infty} z^x \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!} = \mathrm{e}^{-\lambda} \sum_{x=0}^{\infty} \frac{(z\lambda)^x}{x!} = \mathrm{e}^{-\lambda}\mathrm{e}^{z\lambda} = \mathrm{e}^{\lambda(z-1)}.$$

则 mgf 为 $M_X(t) = P_X(\mathrm{e}^t) = \exp[\lambda(\mathrm{e}^t - 1)]$. □

习题

3.20* 某保险业务包含 16 个独立的风险, 每个个体均服从参数为 $\alpha = 1, \theta = 250$ 的 gamma 分布. 利用不完全 gamma 函数给出损失之和超过 6 000 的概率表达式, 再利用中心极限定理近似给出这个值.

3.21* 已知个体赔案的索赔量服从参数为 $\alpha = \frac{8}{3}, \theta = 8\ 000$ 的 Pareto 分布, 利用中心极限定理近似求出 100 个这种独立赔案的总赔付额超过 600 000 的概率.

3.22* 已知个体赔案索赔量服从参数为 $\alpha = 5, \theta = 1\ 000$ 的 gamma 分布 (见附录 A), 利用中心极限定理近似求出 100 个这种独立赔案的总赔付额超过 525 000 的概率.

3.23 由 1 000 份成人健康保险合同构成的样本, 产生了均值为 1 300、标准差为 400 的年均给付保险金. 如果预计明年将签发 2 500 份这样的合同, 利用中心期限定理估计明年的总给付超过期望值 101%的概率.

3.24 证明逆高斯分布 (参数 θ 用 $\beta = \mu^2/\theta$ 代替) 的 mgf 为

$$M(t) = \exp\left[\frac{\mu}{\beta}\left(1 - \sqrt{1-2\beta t}\right)\right], \quad t < 1/(2\beta).$$

第 4 章　分布函数的分类与构造

4.1　引　　言

由分布函数所构成的集合非常大, 很难完全掌握. 当需要为某个随机现象找到适用的分布函数模型时, 我们希望缩小搜寻的区域. 前面讨论过的一种划分是将分布函数划分成离散分布、连续分布和混合分布. 在大多数情况下, 可以明显地分辨出分布函数属于这三类中的哪一类, 然而我们还需要进一步地划分. 在 4.2 节会通过复杂性来划分模型, 4.3 节会根据分布的形状进行区分. 4.4 节会介绍一些构造分布的方法, 之后会列出一些常用的连续分布. 4.5 节是对某些离散分布的扩展应用的介绍. 附录 A 和附录 B 中出现的大部分分布也会列在本章的最后. 虽然本章的主要内容是介绍分布函数之间的区别, 但这些分布也有很多共性, 它们将在精算建模中发挥十分重要的作用. 这些共性包括:

- 分布函数的支集都是非负实数集的子集, 大多数精算现象是计数, 或者对时间或货币的度量, 并且一般是非负的, 尽管当所研究的随机变量为财务上的损益时, 可能会出现负的结果, 然而, 财务损益往往也可以用几个非负变量的和差来表示, 例如表现收入和支出的量;
- 某些分布是更广的一类分布的特例, 这使得人们在建模中, 可以根据复杂程度选择适用的模型;
- 这些分布往往有一个或多个众数, 并且众数之一可能为零.

4.2　参数的作用

本节将探讨这样的问题: 需要多少信息才能详细刻画一个适用的模型. 用来反映信息的某些特征量 (参数) 的个数将在一定程度上代表了模型的复杂程度, 这里所说的一定程度上是指复杂的模型意味着更多的项. 一个简单的模型至少应具有以下特点.

- 当可以用较少的参数确定模型时, 确定每个参数时的精度应该非常高.
- 模型通常要对时间和环境稳定. 也就是说, 如果一个模型在今天适用, 它 (也许会有通货膨胀或类似的现象引起一些微小的变化) 应该在今后或者类似的情形下也适用.
- 因为数据通常是不规则的, 所以简单的模型应该便于进行必要的光滑化处

理.

当然, 复杂的模型也有如下的一些优点.

- 当必须用较多的参数确定模型时, 这个模型与现实尽可能地吻合.
- 当必须用较多的参数确定模型时, 这个模型可以更准确地匹配数据的不规则性.

简单模型常常可以非常精确地进行估计, 但是模型本身可能会过于表面化, 这是另外一种区分模型的方法. 统计建模所遵循的过度节约原则认为应该选择那些能够充分反映现实情况的最简单的模型. 对 "充分" 一词的理解将依赖于模型的用途.

在接下来的小节中, 我们将从简单模型过渡到复杂模型. 对分类的定义存在一些困难, 因为并没有一个一致通用的定义. 除了参数分布外, 其他分布类别将用一些人名代表. 另外这种分类也并不能覆盖所有可能的模型, 也不是每个模型都能很容易地归属于某个类别. 这种分类的讨论可看作是一种定性的分析.

4.2.1 参数分布和尺度分布

这部分的简单模型可以用几个关键的数字来表示.

定义 4.1 *参数分布*(Parametric distribution) *是由分布函数构成的集合, 其中的每个函数由一个或几个特征值所确定, 这些值被称为参数*(parameter). *参数的个数是固定且有限的.*

最熟悉的参数分布是正态分布, 参数为 μ 和 σ^2. 一旦这两个参数给定, 分布函数就可以完全确定.

这些分布是本小节中最简单的分布, 因为通常只需要确定少数几个参数. 附录 A 和 B 中的每个分布都是参数分布. 一般认为, 在这种分布族中, 参数较少的分布比参数较多的分布要简单.

对大多数精算建模工作而言, 如果对随机变量进行数乘变换后不改变分布函数的类别, 这将给建模带来很多的便利. 最常见的这类现象是考虑通货膨胀影响的建模, 以及考虑适用于各种货币单位的模型.

定义 4.2 *称满足如下性质的参数分布构成的集合为尺度分布族*(scale distribution), *当这个分布集合中的任何一个分布所对应的随机变量经过正实数的乘数变换后所生成的随机变量还属于该分布集合.*

例 4.3 *证明指数分布是尺度分布.*

解 根据附录 A, 指数分布的分布函数为 $F_X(x) = 1 - \mathrm{e}^{-x/\theta}$. 令 $Y = cX, c > 0$. 则有

$$F_Y(y) = \Pr(Y \leqslant y) = \Pr(cX \leqslant y) = \Pr\left(X \leqslant \frac{y}{c}\right) = 1 - \mathrm{e}^{-y/c\theta}.$$

这表明 Y 服从参数为 $c\theta$ 的指数分布. □

定义 4.4 对于支集非负的随机变量, 尺度参数(scale parameter) 是满足以下两个条件的尺度分布的一个参数. 首先, 当某个服从尺度分布的随机变量乘上一个正的常数后得到的随机变量的该尺度参数也乘上相同的常数; 其次, 当某个服从尺度分布的随机变量乘上一个正的常数后得到的随机变量的其他的参数都不会改变.

例 4.5 证明 gamma 分布存在尺度参数, 分布的定义见附录 A.

解 设随机变量 X 服从 gamma 分布, 令 $Y = cX, c > 0$, 由附录 A 中不完全 gamma 分布的定义, 有

$$F_Y(y) = \Pr\left(X \leqslant \frac{y}{c}\right) = \Gamma\left(\alpha; \frac{y}{c\theta}\right),$$

这表明 Y 服从参数为 α 和 $c\theta$ 的 gamma 分布. 因此 gamma 分布的参数 θ 是尺度参数. □

很多教科书将 gamma 分布的密度函数表示为

$$f(x) = \frac{x^{\alpha-1}e^{-\beta x}\beta^{\alpha}}{\Gamma(\alpha)}.$$

我们采用了含尺度参数的密度函数形式. 当上面的这种密度函数形式的随机变量乘以 $c(c > 0)$ 时, 参数将变成 α 和 β/c. 同样, 在我们的表达式下均值与 θ 成比例, 而在上述转换形式下均值与 $1/\beta$ 成比例. 在我们的形式下易于直接得到这个参数的估计, 而在转换形式下, 只知道参数与均值的估计是成反比的.

通常也可以通过对分布或密度函数的简单观察识别出尺度参数, 特别是当分布函数中的 x 可以用 x/θ 替换时, 自然可以得到参数 θ.

4.2.2 参数分布族

一些稍复杂的参数分布的形式中参数的个数是有限的, 但不是事先固定的.

定义 4.6 参数分布族(parametric distribution family) 是指由某些参数分布组成的集合, 这些分布之间存在某种关联关系.

下例给出了最普通的一类参数分布族.

例 4.7 考虑基于某个特定的参数分布生成的参数分布族. 该分布族中的其他分布是通过将该特定分布的一个或几个参数取值为一些特殊的数值或令若干个参数相等而得到的. 试证明附录 A 中的变换 beta 分布 (transformed beta, T-Beta) 族是参数分布族.

解 T-Beta 分布有 4 个参数. 分布族中的各种分布都是令 T-Beta 分布中的一些参数等于 1(例如, Pareto 分布相对于参数 $\gamma = \tau = 1$) 或令参数相等 (Paralogistic 分布相当于参数 $\tau = 1$、$\gamma = \alpha$) 后得到的. 其中参数的个数 (2 个到 4 个) 并不是事先确定的, 这使得在定义上有些精细的差异. 有些人在应用 T-Beta 分布时会在可能的取值范围内考虑所有 4 个参数, 而有些人只考虑某些应用特例的可能性, 如

Burr 分布. 例如, 对数据进行初步分析得到 $\tau = 1.01$ 并初步考虑采用这一数值, 但是考虑到 $\tau = 1$ 时为 Burr 分布, 则可能用其进行代替[①]. □

4.2.3 有限混合分布

混合分布本身并不复杂, 在接下来的讨论中我们将考虑用一些方法来增加其复杂性. 人们考虑混合的动机之一是我们所观察的现象可能实际上是具有未知概率的几种现象的混合. 例如, 对于随机抽取的牙医索赔案, 它可能是因为牙齿检查、填充、补牙 (如牙冠) 或外科手术等造成的. 因为以上事件存在不同的发生可能性, 所以有不同的模型, 这时采用混合模型比较恰当.

定义 4.8 称随机变量 Y 是由随机变量 $X_1, \cdots, X_k$ 生成的 k 元混合分布[②](k-point mixture distribution), 如果它的累积分布函数可表示为

$$F_Y(y) = a_1 F_{X_1}(y) + a_2 F_{X_2}(y) + \cdots + a_k F_{X_k}(y), \tag{4.1}$$

其中, 对所有 j 有 $a_j > 0$, 而且 $a_1 + a_2 + \cdots + a_k = 1$ 成立.

这里实际上将 a_j 看作 Y 的实现为随机变量 X_j 的概率. 值得注意的是, 如果某个随机变量有 20 种不同的分布选择, 由其生成的二元混合分布就有超过 200 种[③], 这样对于大多数建模已经足够了. 尽管如此, 仍然存在一些含更多参数的参数分布.

例 4.9 美国保险服务局 (简称 ISO) 的精算师发现一般责任保险的损失模型采用 2 个 Pareto 分布非常成功. 他们还发现不必要考虑 5 个参数, 他们选择的累积分布函数形式为

$$F(x) = 1 - a\left(\frac{\theta_1}{\theta_1 + x}\right)^{\alpha} - (1 - a)\left(\frac{\theta_2}{\theta_2 + x}\right)^{\alpha+2}.$$

值得注意的是, 2 个 Pareto 分布的形状参数相差 2. 第二个分布在取值很小的部分有较大的概率, 它将表示频繁发生且索赔金额较小的模型, 而第一个分布表示索赔金额较大但发生不频繁的模型. 这个分布只有 4 个参数, 一定程度上简化了建模工作.

如果我们事先不知道用于混合分布的个数, 那么 k 也成为一个参数, 下面的定义将说明这一点.

定义 4.10 一般混合分布(变元数不确定)(variable-component mixture distribution) 的分布函数形式如下

---

① 原著在这里并没有直接证明 T-Beta 分布为参数分布, 而是在讨论 T-Beta 分布的性质和特点.
—— 译者注

② "分布的混合" 和 "混合分布" 的含义与半连续半离散分布的情形可以相互替换, 本书不加以区分. 但在正文中会指明讨论的是哪一种分布.

③ 总共有 $\binom{20}{2} + 20 = 210$ 种选择. 附加的 20 个表示出现分布类型相同但参数不同的情形.

$$F(x)=\sum_{j=1}^{K}a_jF_j(x),\sum_{j=1}^{K}a_j=1,\quad a_j>0,\quad j=1,\cdots,K,\quad K=1,2,\cdots.$$

这样的模型称为半参数的, 因为在复杂性上它介于参数模型和非参数模型之间 (见 4.2.4 节). 第 13 章将讨论模型选择问题, 上述模型之间的区别是非常重要的. 当基于数据对参数个数进行估计时, 用来确定合适参数个数的假设检验将变得更加困难. 如果混合分布中所有的变元满足同样的参数分布 (参数不同), 则生成的分布称为“g 变量一般混合”分布, 这里 g 代表每个变元的分布名称.

例 4.11 试给出指数分布的一般混合分布的分布函数、密度函数和损失率函数.

解 指数分布的混合分布函数可以表示为

$$\begin{aligned}F(x)=&1-a_1\mathrm{e}^{-x/\theta_1}-a_2\mathrm{e}^{-x/\theta_2}-\cdots-a_K\mathrm{e}^{-x/\theta_K},\\ \sum_{j=1}^{K}a_j=&1,\quad a_j,\theta_j>0,\quad j=1,\cdots,K,\quad K=1,2,\cdots.\end{aligned}$$

进而有另外两个函数的表示

$$\begin{aligned}f(x)=&a_1\theta_1^{-1}\mathrm{e}^{-x/\theta_1}+a_2\theta_2^{-1}\mathrm{e}^{-x/\theta_2}+\cdots+a_K\theta_K^{-1}\mathrm{e}^{-x/\theta_K},\\ h(x)=&\frac{a_1\theta_1^{-1}\mathrm{e}^{-x/\theta_1}+a_2\theta_2^{-1}\mathrm{e}^{-x/\theta_2}+\cdots+a_K\theta_K^{-1}\mathrm{e}^{-x/\theta_K}}{a_1\mathrm{e}^{-x/\theta_1}+a_2\mathrm{e}^{-x/\theta_2}+\cdots+a_K\mathrm{e}^{-x/\theta_K}}.\end{aligned}$$

这时参数的个数是不确定的甚至可能无限. 例如, 当 $K=2$ 时, 参数为 (a_1,θ_1,θ_2), 这时 a_2 不是一个独立的参数而是被 a_1 唯一确定. 而当 $K=4$ 时就会有 7 个参数. □

例 4.12 举例说明如何用二元混合 gamma 变量构造双峰分布.

解 考虑由两个 gamma 分布的平均构成的混合分布. 其中一个参数是 $\alpha=4,\theta=7$(对应的峰值为 21), 另一个参数是 $\alpha=15,\theta=7$(对应的峰值为 98). 混合分布的密度函数为

$$f(x)=0.5\frac{x^3\mathrm{e}^{-x/7}}{3!7^4}+0.5\frac{x^{14}\mathrm{e}^{-x/7}}{14!7^{15}}$$

密度函数的图像如图 4-1. □

4.2.4 数据依赖型分布

模型 1~5 以及其他的很多情形都依赖于所考虑的现象 (随机变量) 本身的特点, 在对现象进行观测前, 无法确定这些特点. 例如, 在得到牙医的索赔数据前, 完全可以任意假设其分布服从参数为 $\mu=5,\sigma=1$ 的对数正态分布. 但是, 这个模型也许并不能对牙医的索赔进行很好的描述, 不过那是另一个问题. 从另一方面看, 也可以通过数据直接建立模型, 这些模型也会含有参数但通常称其为非参数模型.

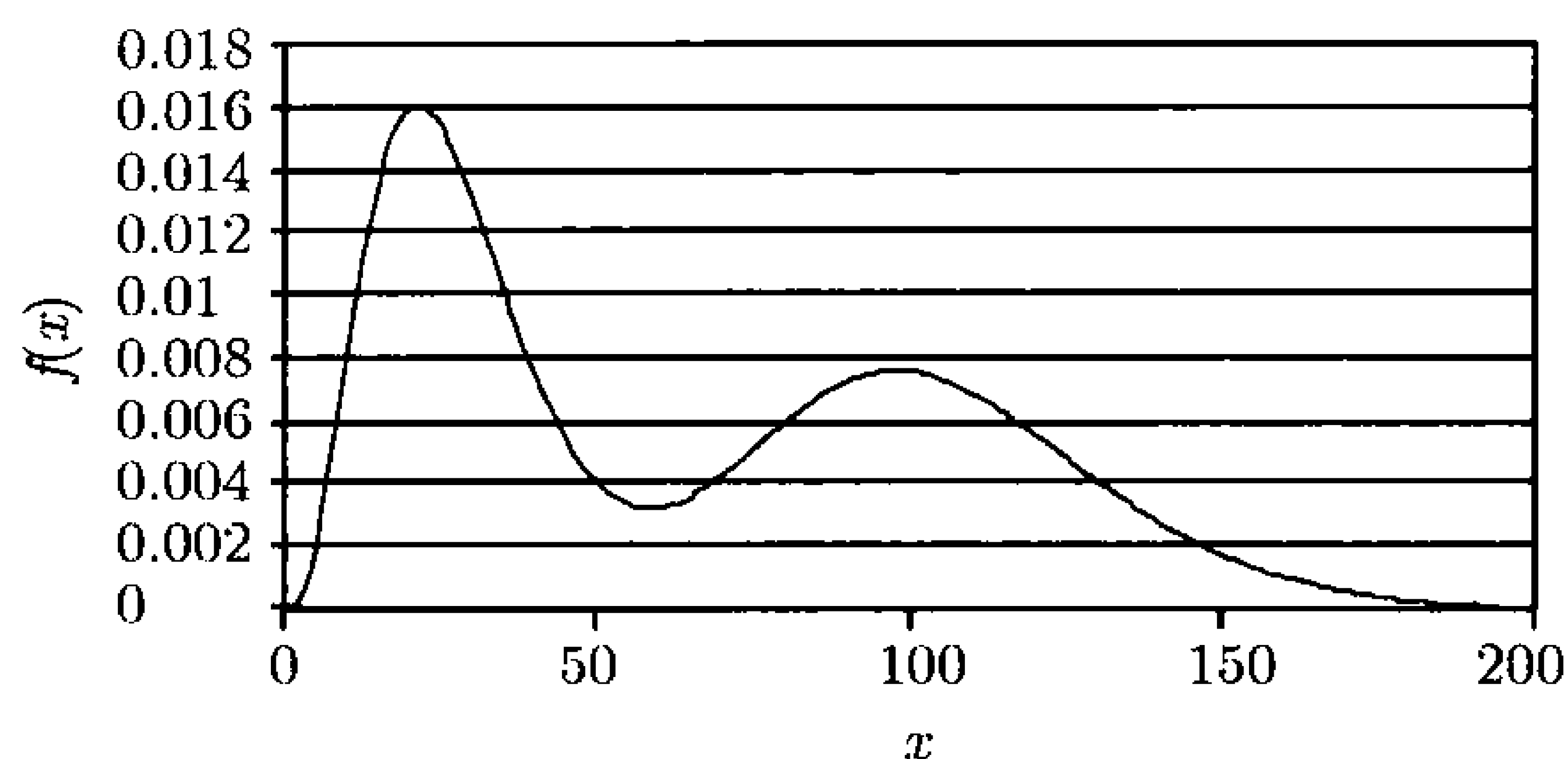


图 4-1 二元混合 gamma 分布

定义 4.13 数据依赖型分布(data-dependent distribution) 是其复杂程度至少不低于其提供的源信息或数据, 并且"参数"的个数随其源信息或数据的增加而增加的一种分布.

本质上, 这些模型的参数个数要多于数据集合的观测值个数. 前面模型 6 的经验分布函数就是一种数据依赖型分布. 概率函数在每个数据点的取值都是 $1/n$, 因此该模型的 n 个参数就是这个生成经验分布的数据集的 n 个观测.

数据依赖型分布的另一个例子是核光滑分布模型, 11.3 节中将会有更详细的介绍. 其模型并非对每个数据点赋予 $1/n$ 的概率, 而是用一个连续函数来取代这个密度函数, 这个连续函数将构筑每个包含数据点的一个区域, 该区域的概率为 $1/n$. 每个区域以数据点为中心, 因此模型与数据相符, 但并不是完美的拟和. 与经验分布相比, 这个模型具有更好的光滑性. 下面给出一个简单的例子.

例 4.14 由模型 6 构造一个核光滑分布模型, 其核函数为均匀分布, 且窗宽为 2.

解 题目所求的概率密度函数可以为

$$f(x)=\sum_{j=1}^{5}p_6(x_j)K_j(x),$$

$$K_j(x)=\begin{cases}0, & |x-x_j|>2,\\ 0.25, & |x-x_j|\leqslant 2,\end{cases}$$

其中的求和是对原模型中有正概率的 5 个点的和. 例如, 求和式中的第一项为

$$p_6(x_1)K_1(x)=\begin{cases}0, & x<1,\\ 0.031\,25, & 1\leqslant x\leqslant 5,\\ 0, & x>5.\end{cases}$$

最终的密度函数是 5 个这类函数的总和, 其图像见图 4-2. □

值得注意的是, 核光滑分布和经验分布都可以表示为混合分布的形式. 这里将

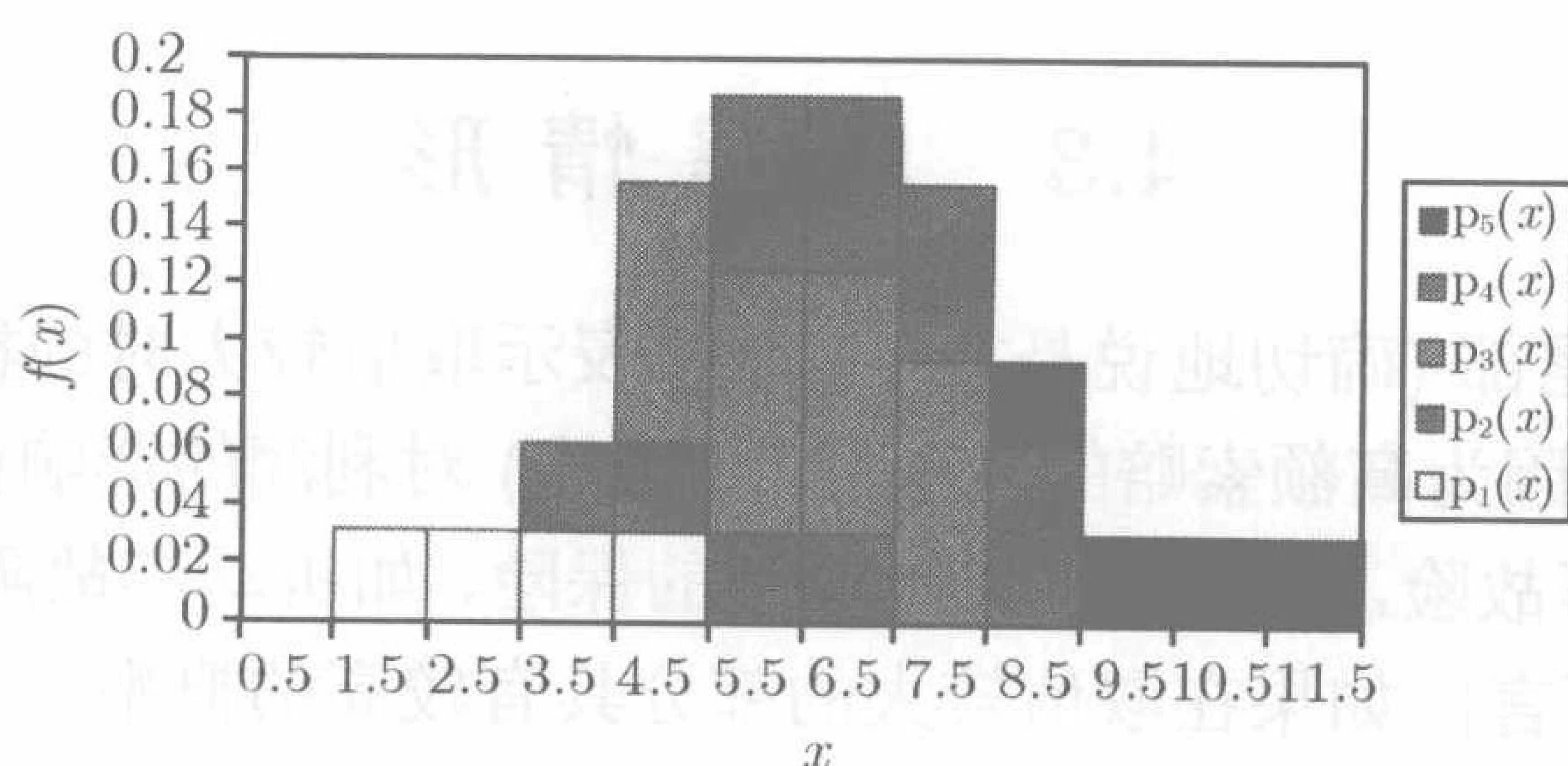


图 4-2 核密度分布

它们单独分类是因为变元的个数与样本规模相关, 而并非与现象本身或是其变量相关.

习题

4.1 证明附录 A 中的对数正态分布为尺度分布, 但并没有尺度参数. 给出该分布的其他参数表示法, 在这种表示下存在尺度参数.

4.2 指出模型 1~6 中属于参数分布的模型, 并描述或指出分布的名称.

4.3* 索赔额服从 $\alpha = 2$ 的 Pareto 分布, θ 未知. 在随后几年中, 索赔额以 6%的速度连续增长. 令 r 表示下一年索赔额超过 d 的部分与本年索赔额超过 d 的部分之比. 当 d 趋于无穷大时, 给出 r 的极限.

4.4 试确定例 4.9 中的二元混合分布的均值和二阶矩. 本题的结果给出了混合分布原始矩的一般公式.

4.5 试确定混合 gamma 分布的均值和方差的表达式.

4.6 模型 1~6 哪些为参数分布族? 哪些为一般混合分布?

4.7* 已知某索赔的 75%服从均值 3 000, 方差 4 000 的正态分布. 其余的 25%服从均值 4 000, 方差 10 000 000 的正态分布. 随机抽取一个样本, 试确定其索赔超过 5 000 的概率.

4.8* X 服从参数为 $\alpha = 1, \gamma = 2, \theta = \sqrt{1\ 000}$ 的 Burr 分布, Y 服从参数为 $\alpha = 1, \theta = 1\ 000$ 的 Pareto 分布. Z 是由 X 和 Y 平均加权产生的混合分布. 试确定 Z 的中位数. 令 W=1.1Z. 证明 W 也是由 Burr 分布和 Pareto 分布混合而成, 并确定 W 的参数.

4.9* 考虑三个随机变量: X 由 (0,2) 上的均匀分布和 (0,3) 上的均匀分布混合而成. Y 是两个随机变量的和, 两个随机变量分别服从 (0,2) 上的均匀分布和 (0,3) 上的均匀分布. Z 是正态分布的随机变量在 1 这个点右删失. 从下列叙述中找出符合上述描述的变量:

(a) 分布函数和密度函数都连续;

(b) 密度函数连续, 分布函数不连续;

(c) 分布函数不连续.

4.10 证明例 4.14 中的模型是混合均匀分布.

4.11 证明附录 A 中列出的逆高斯分布属于尺度分布族, 但不存在尺度参数.

4.12 证明 Weibull 分布存在尺度参数.

4.3 厚尾情形

分布密度的尾部 (确切地说是右尾) 通常表示取值较大时的概率. 精算师对这个部分非常关注, 因为高额索赔的发生 (或不发生) 对利润的影响很大. 高风险类型的保险，如医疗事故险，将会比低风险类型的保险，如机动车故障险，具有更大的索赔 (相对均值而言). 如果在取值较大的部分具有较高的概率, 一般称这种随机变量具有厚尾. 厚尾是一个相对的概念 (模型 A 比模型 B 的尾部更厚) 或是一个绝对的概念 (具有某种特征的分布被称为厚尾分布). 当选择模型时, 考虑尾部的厚度可以帮助我们缩小选择的范围. 例如, 对医疗事故索赔考虑 Pareto 分布模型是较为合理的, 因为 Pareto 分布的尾部很厚. 相反地, 对牙医保险则应考虑对数正态分布模型, 因为其尾部很轻. 然而, 还应该注意的是测量尾部轻重程度的方法并不一致.

4.3.1 矩的存在性

取正值的连续随机变量 (正如大多保险赔付变量) 的 k 阶原点矩定义为 $\int_0^\infty x^k f(x)\mathrm{d}x$, 这个积分结果依赖于密度函数和 k, 当然也可能积分不存在. 如果在 x 很大时密度函数也很大, 再乘上 x^k, 有可能导致积分不收敛. 因此, 如果任意的正数阶原点矩存在将说明分布的尾部很轻, 而正数阶矩存在最高阶数 (或不存在正数阶矩) 则说明分布的尾部很厚①.

例 4.15　*证明 gamma 分布的所有正数阶矩都存在, 而 Pareto 分布并非所有的正数阶矩都存在.*

证明　对于 gamma 分布, 有

$$\begin{aligned}\mu_k' &= \int_0^\infty x^k \frac{x^{\alpha-1}\mathrm{e}^{-x/\theta}}{\Gamma(\alpha)\theta^\alpha}\mathrm{d}x \\ &= \int_0^\infty (y\theta)^k \frac{(y\theta)^{\alpha-1}\mathrm{e}^{-y}}{\Gamma(\alpha)\theta^\alpha}\theta\mathrm{d}y \quad \text{考虑替换：} y = x/\theta\end{aligned}$$

① 同样, 任意负数阶矩的存在性说明尾部很轻. 左边尾部的情况也会帮助我们选择合适的模型. 下面的讨论适用于零附近单调可微的密度函数. 特别地, $f(0)$ 以及密度函数在零附近的斜率都将影响负数阶矩 $\mathrm{E}(X^k)$(k 为负数) 的存在性. 假设负数阶矩仅在 $k > -r$ 时存在. 若 $r < 1$, 则当 $x \to 0$ 时, $f(x)$ 趋于无穷. 若 $r = 1$, $f(0)$ 是一个非负数. 若 $1 < r < 2$, $f(0) = 0$ 且, 当 $x \to 0$ 时斜率趋于无穷. 若 $r = 2$, $f(0) = 0$ 并且函数在零点的斜率为非负数. 若 $r > 2$, $f(0) = 0$ 并且函数在零点的斜率为 0, 所以分布在零点附近很少. 考虑如下的例子: 参数 $\tau = 0.2$ 的 Weibull 分布的工伤保险模型, $r = 0.2$, 说明函数在 0 附近的概率很大. 确定 θ 使其均值为 30 000, 进而得到索赔小于 1 的概率为 28%, 超过 500 000 的概率为 1%. 这是一个大额损失的恰当模型. (参见 4.4.7 节分段模型的分法). 相比之下, 对数正态分布的均值和方差相等, 其小于 1 的概率低于 0.1%, 超过 500 000 的概率约为 1%. 对数正态分布的所有负数阶矩都存在, 因此当 x 趋于 0 时 $f(x)$ 趋于 0.

$$= \frac{\theta^k}{\Gamma(\alpha)}\Gamma(\alpha+k) < \infty, \quad 对所有k > 0,$$

而对于 Pareto 分布, 有

$$\begin{aligned}
\mu'_k &= \int_0^\infty x^k \frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}}\mathrm{d}x \\
&= \int_0^\infty (y-\theta)^k \frac{\alpha\theta^\alpha}{y^{\alpha+1}}\mathrm{d}y \quad 考虑替换: y = x+\theta \\
&= \alpha\theta^\alpha \int_\theta^\infty \sum_{j=0}^{k} \binom{k}{j} y^{j-\alpha-1}(-\theta)^{k-j}\mathrm{d}y, \quad 对所有整数k.
\end{aligned}$$

上述积分存在仅当和式中 y 的指数都小于 -1. 也就是对所有满足 $j-\alpha-1<-1$ 的 j 都成立, 或等价于 $k<\alpha$. 因此, Pareto 分布只有某些矩存在. □

由本例看出, Pareto 分布比 gamma 分布尾部更厚. 由附录 A 中的矩表达式可以看出哪些分布的尾部更厚, 这也可以由矩的存在性推出的.

4.3.2 极限比

对两个分布尾部厚度的比较也可以通过计算两个生存函数的比值在趋于无穷时的收敛性 (尾部较厚的分布作为分子) 来进行. 如果这个比值收敛, 则表明处于分子的分布函数在数值较大处明显有更大的概率值. 这等价于计算密度函数的比率. 极限将是相同的, 因为由 L'Hopital(洛必达) 法则, 有

$$\lim_{x\to\infty}\frac{S_1(x)}{S_2(x)} = \lim_{x\to\infty}\frac{S'_1(x)}{S'_2(x)} = \lim_{x\to\infty}\frac{-f_1(x)}{-f_2(x)}.$$

例 4.16 利用密度函数比率的极限证明 Pareto 分布比 gamma 分布尾部更厚.

证明 为避免混淆, 将 gamma 分布常用的参数 α 和 θ 用 τ 和 λ 表示. 所求极限为

$$\begin{aligned}
\lim_{x\to\infty}\frac{f_{\text{Pareto}}(x)}{f_{\text{gamma}}(x)} &= \lim_{x\to\infty}\frac{\alpha\theta^\alpha(x+\theta)^{-\alpha-1}}{x^{\tau-1}\mathrm{e}^{-x/\lambda}\lambda^{-\tau}\Gamma(\tau)^{-1}} \\
&= c\lim_{x\to\infty}\frac{\mathrm{e}^{x/\lambda}}{(x+\theta)^{\alpha+1}x^{\tau-1}} > c\lim_{x\to\infty}\frac{\mathrm{e}^{x/\lambda}}{(x+\theta)^{\alpha+\tau}}.
\end{aligned}$$

或者利用 L'Hôpital 法则, 或者利用指数比多项式趋于无穷大的速度更快这个性质, 可以得到上述极限为无穷大. 图 4-3 给出了 Pareto 分布 (参数 α=3, θ=10) 和 gamma 分布 (参数 α=1/3, θ=15) 的部分密度函数. 两个分布都满足均值为 5, 方差为 75. 图 4-3 的表现与上面进行的代数推导结果是一致的. □

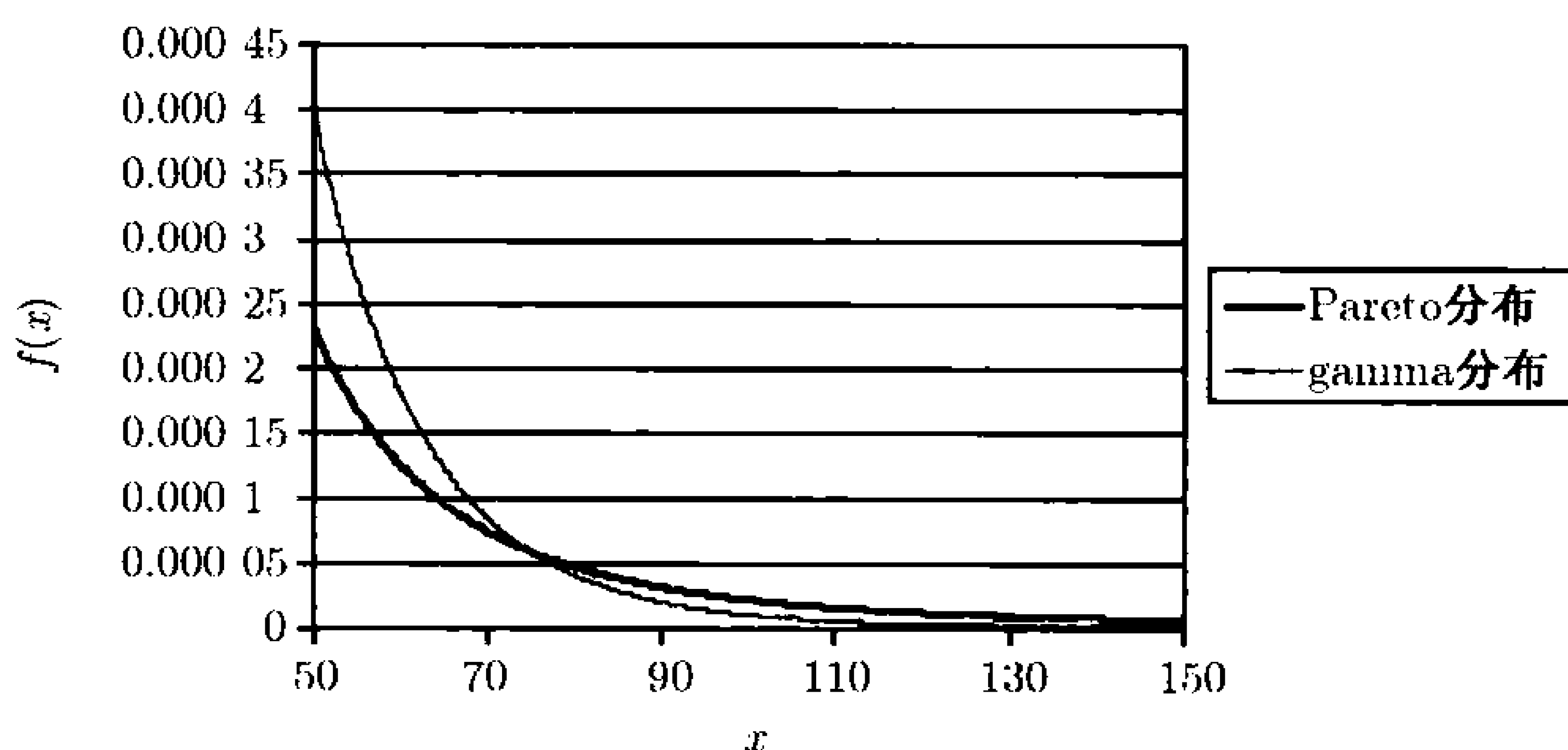


图 4-3　gamma 分布和 Pareto 分布的尾部

4.3.3　损失率和平均剩余生命函数

损失率函数的性质也可以给出一些尾部分布的信息. 如果损失率函数递减, 则在数值不大处的概率变小而在数值较大处的概率变大. 因此分布的尾部较厚. 相反, 如果损失率函数递增, 则说明尾部较轻.

例 4.17　利用损失率函数比较 Pareto 分布和 gamma 分布的尾部.

解　Pareto 分布的损失率函数为

$$h(x)=\frac{f(x)}{S(x)}=\frac{\alpha\theta^{\alpha}(x+\theta)^{-\alpha-1}}{\theta^{\alpha}(x+\theta)^{-\alpha}}=\frac{\alpha}{x+\theta},$$

它是递减函数. 对于 gamma 分布, 因为无法给出 $S(x)$ 的解析表达式, 需要一些推导, 观察到

$$\frac{1}{h(x)}=\frac{\int_x^{\infty}f(t)\mathrm{d}t}{f(x)}=\frac{\int_0^{\infty}f(x+y)\mathrm{d}y}{f(x)}.$$

所以, 对于固定的 y, 如果 $f(x+y)/f(x)$ 是关于 x 的增函数, 则 $1/h(x)$ 也是关于 x 递增, 因此损失率函数为递减的. 而对于 gamma 分布, 有

$$\frac{f(x+y)}{f(x)}=\frac{(x+y)^{\alpha-1}\mathrm{e}^{-(x+y)/\theta}}{x^{\alpha-1}\mathrm{e}^{-x/\theta}}=\left(1+\frac{y}{x}\right)^{\alpha-1}\mathrm{e}^{-y/\theta},$$

这个比值在 $\alpha<1$ 时关于 x 严格递增, 在 $\alpha>1$ 时严格递减. 利用这种方法可以知道, 有些 gamma 分布 ($\alpha<1$) 的尾部较厚, 有些 gamma 分布的尾部较轻. 当 $\alpha=1$ 时, 为指数分布, 损失率为常数. 尽管 gamma 函数的 $h(x)$ 比较复杂, 我们还是可以知道当 x 很大时的情况. 由于当 $x\to\infty$ 时 $f(x)$ 和 $S(x)$ 都趋于 0, 由 L'Hôpital 法则:

$$\begin{aligned}\lim_{x\to\infty}h(x)&=\lim_{x\to\infty}\frac{f(x)}{S(x)}=-\lim_{x\to\infty}\frac{f'(x)}{f(x)}=-\lim_{x\to\infty}\left[\frac{\mathrm{d}}{\mathrm{d}x}\ln f(x)\right]\\&=-\lim_{x\to\infty}\frac{\mathrm{d}}{\mathrm{d}x}\left[(\alpha-1)\ln x-\frac{x}{\theta}\right]=\lim_{x\to\infty}\left(\frac{1}{\theta}-\frac{\alpha-1}{x}\right)=\frac{1}{\theta}.\end{aligned}$$

因此, 当 $x \to \infty$ 时, $h(x) \to 1/\theta$. □

平均剩余生命函数也能表现关于厚尾的信息. 如果平均剩余生命函数随 d 递增, 那么在变量取值较大处的期望结果会很大, 因此概率向右移, 说明其尾部相比那些平均剩余生命函数递减或增速较慢的模型更厚. 事实上, 平均剩余生命函数和损失率在很多方面都紧密相关. 首先, 有

$$\begin{aligned}\frac{S(y+d)}{S(d)} &= \frac{\exp\left[-\int_0^{y+d} h(x)\mathrm{d}x\right]}{\exp\left[-\int_0^{d} h(x)\mathrm{d}x\right]} = \exp\left[-\int_d^{y+d} h(x)\mathrm{d}x\right] \\ &= \exp\left[-\int_0^{y} h(d+t)\mathrm{d}t\right].\end{aligned}$$

因此, 如果损失率递减, 则对于固定的 y, $\int_0^y h(d+t)\mathrm{d}t$ 为 d 的减函数, 从上面可以知道 $S(y+d)/S(d)$ 为 d 的增函数. 而由 (3.5), 平均剩余生命函数可以表示成

$$e(d) = \frac{\int_d^\infty S(x)\mathrm{d}x}{S(d)} = \int_0^\infty \frac{S(y+d)}{S(d)}\mathrm{d}y.$$

因此, 如果损失率递减, 可以知道平均剩余生命函数 $e(d)$ 为 d 的增函数, 因为这个结论对于固定的 y 关于 $S(y+d)/S(d)$ 也是成立的. 类似的, 如果损失率函数为递增的, 则平均剩余生命函数为 d 的减函数. 然而, 值得注意的是 (或许有悖直觉) 反过来的推理并不成立. 习题 4.16 给出一个例子, 其平均剩余生命函数递减, 而损失率却处处不增. 然而, 上面所述的推理一般和尾部厚度的讨论一致.

关于平均剩余生命函数和损失率还有一层关系. 当 $d \to \infty$ 时, $S(d)$ 和 $\int_d^\infty S(x)\mathrm{d}x$ 都趋于 0. 因此, 由 (3.5) 式, 当 $d \to \infty$ 时平均剩余生命函数的极限可以由 L'Hôpital 法则确定. 只要所示的极限存在, 有

$$\lim_{d\to\infty} e(d) = \lim_{d\to\infty} \frac{\int_d^\infty S(x)\mathrm{d}x}{S(d)} = \lim_{d\to\infty} \frac{-S(d)}{-f(d)} = \lim_{d\to\infty} \frac{1}{h(d)}.$$

当 $S(x)$(因此可知 $h(x)$ 和 $e(d)$) 非常复杂时, 这些极限间的关系会对分析有帮助.

例 4.18 验证 gamma 分布的平均剩余生命函数的特点.

解 因为 $e(d) = \int_d^\infty S(x)\mathrm{d}x/S(d)$ 并且 $S(x)$ 非常复杂, 因此 $e(d)$ 也很复杂. 但由附录 A 可得 $e(0) = \mathrm{E}(X) = \alpha\theta$, 再利用例 4.17, 有

$$\lim_{x\to\infty} e(x) = \lim_{x\to\infty} \frac{1}{h(x)} = \frac{1}{\lim\limits_{x\to\infty} h(x)} = \theta.$$

同样由例 4.17 知, $\alpha < 1$ 时 $h(x)$ 关于 x 严格递减, $\alpha > 1$ 时 $h(x)$ 关于 x 严格递增, 这意味着: 当 $\alpha < 1$ 时, $e(d)$ 由 $e(0) = \alpha\theta$ 严格递增至 $e(\infty) = \theta$; 当 $\alpha > 1$

时, 而 $e(d)$ 由 $e(0)=\alpha\theta$ 严格递减至 $e(\infty)=\theta$; 当 $\alpha=1$ 时, 分布为指数分布, $e(d)=\theta$. □

要对平均剩余生命函数和尾部厚度进一步深入研究, 需要引入所谓的均衡分布, 这个概念在第 8 章的连续时间破产模型中有着重要的应用. 对于一个正值随机变量, 且 $S(0)=1$, 由定义 3.6 和 (3.5) 式令 $d=0$, 有 $\mathrm{E}(X)=\int_0^\infty S(x)\mathrm{d}x$, 等价于 $1=\int_0^\infty S(x)\mathrm{d}x/E(X)$ 成立, 所以, 可以将

$$f_e(x)=\frac{S(x)}{\mathrm{E}(X)},\quad x\geqslant 0, \tag{4.2}$$

看作为一个新的概率密度函数 (也就是均衡分布), 其对应的生存函数为

$$S_e(x)=\int_x^\infty f_e(t)\mathrm{d}t=\frac{\int_x^\infty S(t)\mathrm{d}t}{\mathrm{E}(X)},\quad x\geqslant 0.$$

利用 (3.5), 这个均衡分布的损失率为

$$h_e(x)=\frac{f_e(x)}{S_e(x)}=\frac{S(x)}{\int_x^\infty S(t)\mathrm{d}t}=\frac{1}{e(x)}.$$

因此平均剩余生命函数的倒数是该均衡分布的损失率函数, 这样的关系可以用来证明平均剩余生命函数可以唯一确定原来的分布. 我们有

$$f_e(x)=h_e(x)S_e(x)=h_e(x)\mathrm{e}^{-\int_0^x h_e(t)\mathrm{d}t},$$

或等价地

$$S(x)=\frac{e(0)}{e(x)}\mathrm{e}^{-\int_0^x\left\{\frac{1}{e(t)}\right\}\mathrm{d}t},$$

这一步将用到 $e(0)=\mathrm{E}(X)$.

均衡分布也可以进一步反映损失率、平均剩余生命函数和尾部厚度之间的关系. 假设 $S(0)=1$, 因此 $e(0)=\mathrm{E}(X)$, 得到 $\int_x^\infty S(t)\mathrm{d}t=e(0)S_e(x)$ 并且由 (3.5) 式, $\int_x^\infty S(t)\mathrm{d}t=e(x)S(x)$. 联立两个等式得到

$$\frac{e(x)}{e(0)}=\frac{S_e(x)}{S(x)}.$$

如果平均剩余生命函数是递增的 (可以导出损失率递减), 则 $e(x)\geqslant e(0)$, 由上面的等式这显然等价于 $S_e(x)\geqslant S(x)$. 进而可以推出

$$\int_0^\infty S_e(x)\mathrm{d}x\geqslant\int_0^\infty S(x)\mathrm{d}x.$$

由定义 3.6 和 (3.5) 式知, 若 $S(0)=1$ 则有 $\mathrm{E}(X)=\int_0^\infty S(x)\mathrm{d}x$. 同样, 因为两边都是均衡分布均值的表达式, $\int_0^\infty S_e(x)\mathrm{d}x=\int_0^\infty xf_e(x)\mathrm{d}x$ 成立. 在 (3.9) 式中令 $u=\infty$, $k=2$ 和 $F(0)=0$ 就可以给出均衡分布的均值, 即

$$\int_0^\infty S_e(x)\mathrm{d}x=\int_0^\infty xf_e(x)\mathrm{d}x=\frac{1}{\mathrm{E}(X)}\int_0^\infty xS(x)\mathrm{d}x=\frac{\mathrm{E}(X^2)}{2\mathrm{E}(X)}.$$

因此有以下不等式

$$\frac{\mathrm{E}(X^2)}{2\mathrm{E}(X)}\geqslant \mathrm{E}(X),$$

或利用 $\mathrm{Var}(X)=\mathrm{E}(X^2)-\mathrm{E}(X)^2$ 有 $\mathrm{Var}(X)\geqslant \mathrm{E}(X)^2$. 也就是说, 若 $e(x)\geqslant e(0)$, 变异系数的平方或其本身大于等于 1. 若 $e(x)\leqslant e(0)$, 不等式要颠倒过来且变异系数小于等于 1, 反过来意味着平均剩余生命函数递减或是损失率函数递增. 这些变异系数的值与这里讨论的尾部厚度是一致的.

习题

4.13 用本节中的方法 (除去平均剩余生命) 比较 Weibull 分布和逆 Weibull 分布的尾部.

4.14 例 4.16 中指出对数正态分布的尾部介于 gamma 分布和 Pareto 分布之间. 为了加强这个结论, 考虑参数 α=0.2, θ=500 的 gamma 分布; 参数 μ=3.709 290, σ=1.338 566 的对数正态分布; 参数为 α=2.5, θ=150 的 Pareto 分布. 首先证明三个分布的均值和方差相同. 其次证明存在某个常数, 使得位于此数值以上的任意点, gamma 分布的概率分布函数都小于对数正态概率分布函数和 Pareto 概率分布函数; 并存在另外的常数, 使得位于此数值以上的任意点, 对数正态分布的概率分布函数都小于 gamma 概率分布函数和 Pareto 概率分布函数.

4.15 随机变量 Y 为 (4.2) 中的均衡分布. 即 $f_Y(y)=f_e(y)=S_X(y)/\mathrm{E}(X)$, 对某个随机变量 X 成立. 若 $M_X(t)$ 存在, 利用分部积分证明

$$M_Y(t)=\frac{M_X(t)-1}{t\mathrm{E}(X)}.$$

4.16 随机变量 X 的概率密度函数为 $f(x)=(1+2x^2)\mathrm{e}^{-2x}, x\geqslant 0$.

(a) 确定生存函数 $S(x)$.

(b) 确定损失率函数 $h(x)$.

(c) 确定均衡分布的生存函数 $S_e(x)$.

(d) 确定平均剩余生命函数 $e(x)$.

(e) 确定 $\lim\limits_{x\to\infty} h(x)$, $\lim\limits_{x\to\infty} e(x)$.

(f) 证明 $e(x)$ 严格递减, 而 $h(x)$ 并非严格递增.

4.17 假设 X 的概率密度函数为 $f(x)$, $x\geqslant 0$.

(a) 证明

$$S_e(x)=\frac{\int_x^{\infty}(y-x)f(y)\mathrm{d}y}{\mathrm{E}(X)}.$$

(b) 利用 (a) 证明

$$\int_x^{\infty}yf(y)\mathrm{d}y=xS(x)+\mathrm{E}(X)S_e(x).$$

(c) 证明 (b) 可以写成

$$S(x)=\frac{\int_x^{\infty}yf(y)\mathrm{d}y}{x+e(x)},$$

还可以推出

$$S(x)\leqslant\frac{\mathrm{E}(X)}{x+e(x)}.$$

(d) 利用 (c) 证明, 如果 $e(x)\geqslant e(0)$, 则有

$$S(x)\leqslant\frac{\mathrm{E}(X)}{x+\mathrm{E}(X)},$$

并且说明其均值至少等于 (最小的) 中位数.

(e) 证明 (b) 也可以表示为

$$S_e(x)=\frac{e(x)}{x+e(x)}\cdot\frac{\int_x^{\infty}yf(y)\mathrm{d}y}{\mathrm{E}(X)},$$

因此

$$S_e(x)\leqslant\frac{e(x)}{x+e(x)}.$$

4.4　构造新的分布

4.4.1　引言

本节将指出如何根据已有的分布构造新的参数分布. 附录 A 中的很多分布都是通过这种方法得到的.

4.4.2　倍数变换

这种变换等价于在各种损失上一致地引入通胀因素, 也就是常说的标度变换. 例如, 年损失随机变量为 X, 通货膨胀率为 5%, 则下一年的损失随机变量可以用 $Y=1.05X$ 来建模.

定理 4.19　X 为连续随机变量, 概率分布函数 $f_X(x)$, 累积分布函数 $F_X(x)$. 令 $Y=\theta X$, $\theta>0$. 则

$$F_Y(y)=F_X\left(\frac{y}{\theta}\right),\quad F_Y(y)=\frac{1}{\theta}f_X\left(\frac{y}{\theta}\right).$$

证明

$$F_Y(y) = \Pr(Y \leqslant y) = \Pr(\theta X \leqslant y) = \Pr\left(X \leqslant \frac{y}{\theta}\right) = F_X\left(\frac{y}{\theta}\right),$$

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}F_Y(y) = \frac{1}{\theta}f_X\left(\frac{y}{\theta}\right).$$

□

推论 4.20 参数 θ 是随机变量 Y 的尺度参数.

下面的例子将说明这个过程.

例 4.21 设 X 的概率密度函数为 $f(x) = \mathrm{e}^{-x}, x > 0$. 确定 $Y = \theta X$ 的累积分布函数和概率密度函数.

解

$$F_X(x) = 1 - \mathrm{e}^{-x}, \quad F_Y(y) = 1 - \mathrm{e}^{-y/\theta}, \quad f_Y(y) = \frac{1}{\theta}\mathrm{e}^{-y/\theta}.$$

可以看出这是参数为 θ 的指数分布. □

4.4.3 幂变换

定理 4.22 设 X 是连续随机变量, 其概率分布函数为 $f_X(x)$, 累积分布函数为 $F_X(x)$, 并且 $F_X(0) = 0$. 令 $Y = X^{1/\tau}$, 则若 $\tau > 0$,

$$F_Y(y) = F_X(y^\tau), \quad f_Y(y) = \tau y^{\tau-1} f_X(y^\tau), \quad y > 0.$$

若 $\tau < 0$,

$$F_Y(y) = 1 - F_X(y^\tau), \quad f_Y(y) = -\tau y^{\tau-1} f_X(y^\tau). \tag{4.3}$$

证明 若 $\tau > 0$,

$$F_Y(y) = \Pr(X \leqslant y^\tau) = F_X(y^\tau),$$

若 $\tau < 0$,

$$F_Y(y) = \Pr(X \geqslant y^\tau) = 1 - F_X(y^\tau).$$

概率分布函数由上式微分得到. □

通常希望参数取正值, 如果 τ 为负, 则构造新参数 $\tau^* = -\tau$. 那么 (4.3) 式变成

$$F_Y(y) = 1 - F_X(y^{-\tau^*}), \quad f_Y(y) = \tau^* y^{-\tau^*-1} f_X(y^{-\tau^*}).$$

以后可以省略星号直接用这个正参数.

定义 4.23 当采用幂函数进行分布的变换时, 如果 $\tau > 0$, 一般称为原分布的变换分布(transformed), 如果 $\tau = -1$, 一般称为原分布的逆分布(inverse), 而如果 $\tau < 0$(但非 -1), 一般称为原分布的逆变换分布(inverse transformed). 为了构造附录 A 中的各种分布, 并使得参数 θ 始终为尺度参数, 应在基本分布乘以 θ 之前先进行幂变换.

例 4.24 假设 X 服从指数分布. 确定其逆分布、变换分布和逆变换分布的累积分布函数.

解 没有尺度参数的逆指数分布的累积分布函数为

$$F(y) = 1 - [1 - \mathrm{e}^{-1/y}] = \mathrm{e}^{-1/y}.$$

加入尺度参数后为 $F(y) = \mathrm{e}^{-\theta/y}$.

没有尺度参数的变换指数分布的累积分布函数为

$$F(y) = 1 - \exp(-y^{\tau}).$$

加入尺度参数后为 $F(y) = 1 - \exp[-(y/\theta)^{\tau}]$, 通常称其为 **Weibull 分布**.

没有尺度参数的逆变换指数分布的累积分布函数为

$$F(y) = 1 - [1 - \exp(-y^{-\tau})] = \exp(-y^{-\tau}).$$

加入尺度参数后为 $F(y) = \exp[-(\theta/y)^{\tau}]$, 为**逆 Weibull 分布**. □

另一个基本分布的概率密度函数是 $f(x) = x^{\alpha-1}\mathrm{e}^{-x}/\Gamma(\alpha)$. 当加入尺度参数后变为 **gamma 分布**. 利用本节的结果也可以构造其逆分布和变换分布, 但与其他分布不同的是, 它并没有累积分布函数的解析表达, 我们最多只能为这类函数进行一些标准化的定义.

定义 4.25 参数为 $\alpha > 0$ 的不完全 gamma 函数定义如下

$$\Gamma(\alpha; x) = \frac{1}{\Gamma(\alpha)} \int_0^x t^{\alpha-1}\mathrm{e}^{-t}\mathrm{d}t,$$

其中 gamma 函数定义如下

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1}\mathrm{e}^{-t}\mathrm{d}t.$$

另外有 $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$, 并且对一切正整数 n 有 $\Gamma(n) = (n-1)!$. 附录 A 给出了这些量的数值计算方法. 在大多数电子表格系统和统计及数值分析程序中都嵌有这些函数的数值.

4.4.4 指数变换

定理 4.26 连续随机变量 X 的概率密度函数和累积分布函数分别为 $f_X(x)$ 和 $F_X(x)$, 其中 $f_X(x) > 0$ 对所有实数 x 成立. 令 $Y = \mathrm{e}^X$, 则对 $y > 0$, 有

$$F_Y(y) = F_X(\ln y), \quad f_Y(y) = \frac{1}{y} f_X(\ln y).$$

证明 $F_Y(y) = \Pr(e^X \leqslant y) = \Pr(X \leqslant \ln y) = F_X(\ln y)$. □

例 4.27 设 X 服从均值为 μ 方差为 σ^2 的正态分布. 确定 $Y = e^X$ 的概率密度函数和累积分布函数.

解

$$F_Y(y) = \Phi\left(\frac{\ln y - \mu}{\sigma}\right),$$

$$f_Y(y) = \frac{1}{y\sigma}\phi\left(\frac{\ln y - \mu}{\sigma}\right) = \frac{1}{y\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\ln y - \mu}{\sigma}\right)^2\right].$$ □

可以试着通过 $W = \theta Y$ 加入一个尺度参数, 但这样没有效果, 证明见习题 4.22. 本例构造出了**对数正态分布**(我们用这样的名称, 尽管"指数正态"可能更合适).

4.4.5 混合

混合的概念可以由有限多个随机变量的混合延伸到无限多个随机变量的混合. 在以下定理中概率分布函数 $f_\Lambda(\lambda)$ 代替了前面 k 元混合中的离散概率 a_j.

定理 4.28 已知 X 的条件概率密度函数为 $f_{X|\Lambda}(x|\lambda)$ 且累积分布函数为 $F_{X|\Lambda}(x|\lambda)$, 这里 λ 是 X 的一个参数. X 可能还有其他的参数, 但都不相关. 随机变量 Λ 的概率密度函数是 $f_\Lambda(\lambda)$, λ 是随机变量 Λ 的实现, 则 X 的非条件概率分布函数为

$$f_X(x) = \int f_{X|\Lambda}(x|\lambda) f\Lambda(\lambda)\mathrm{d}\lambda. \tag{4.4}$$

积分是对所有具有正概率的 λ 值进行的, 这样得到的分布是一个混合分布. 分布函数可以表示如下

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x}\int f_{X|\Lambda}(y|\lambda) f\Lambda(\lambda)\mathrm{d}\lambda\mathrm{d}y = \int\int_{-\infty}^{x} f_{X|\Lambda}(y|\lambda) f\Lambda(\lambda)\mathrm{d}y\mathrm{d}\lambda \\ &= \int F_{X|\Lambda}(X|\lambda) f\Lambda(\lambda)\mathrm{d}\lambda. \end{aligned}$$

混合分布的矩函数为

$$\mathrm{E}(X^k) = \mathrm{E}[\mathrm{E}(X^k|\Lambda)].$$

特别地,

$$\mathrm{Var}(X) = \mathrm{E}[\mathrm{Var}(X|\Lambda)] + \mathrm{Var}[\mathrm{E}(X|\Lambda)].$$

证明 由定义, 被积函数是 X 和 Λ 的联合密度函数. 则积分值为边缘密度函数. 期望值为 (假设积分顺序可以颠倒)

$$\begin{aligned} \mathrm{E}(X^k) &= \int\int x^k f_{X|\Lambda}(x|\lambda) f_\Lambda(\lambda)\mathrm{d}\lambda\mathrm{d}x = \int\left[\int x^k f_{X|\Lambda}(x|\lambda)\mathrm{d}x\right] f_\Lambda(\lambda)\mathrm{d}\lambda \\ &= \int \mathrm{E}(X^k|\lambda) f_\Lambda(\lambda)\mathrm{d}\lambda = \mathrm{E}[\mathrm{E}(X^k|\Lambda)]. \end{aligned}$$

方差为

$$\begin{aligned}\mathrm{Var}(X)&=\mathrm{E}(X^2)-[\mathrm{E}(X)]^2=\mathrm{E}[\mathrm{E}(X^2|\Lambda)]-\{\mathrm{E}[\mathrm{E}(X|\Lambda)]\}^2\\&=\mathrm{E}\{\mathrm{Var}(X|\Lambda)+[\mathrm{E}(X|\Lambda)]^2\}-\{\mathrm{E}[\mathrm{E}(X|\Lambda)]\}^2\\&=\mathrm{E}[\mathrm{Var}(X|\Lambda)]+\mathrm{Var}[\mathrm{E}(X|\Lambda)].\end{aligned}$$

值得注意的是, 如果 $f_\Lambda(\lambda)$ 是离散的, 积分将被求和式替换. 可以将结果写为 $f_X(x)=\mathrm{E}_\Lambda[f_{X|\Lambda}(x|\Lambda)]$, $F_X(x)=\mathrm{E}_\Lambda[F_{X|\Lambda}(x|\Lambda)]$, 带下标的 E 表示这时的随机变量是 Λ. □

一个有趣的现象是混合分布通常尾部较厚, 因此这是一个构造这类模型的好方法. 特别地, 如果 $f_{X|\Lambda}(x|\lambda)$ 存在一个对所有 λ 递减的损失率函数, 则混合分布也有一个递减的损失率函数 (详见 Ross[114]pp.407~409). 下例将利用混合方法构造一个尾部较厚的常见分布.

例 4.29 设 $X|\Lambda$ 服从参数为 $1/\Lambda$ 的指数分布, 并且设 Λ 服从 gamma 分布. 确定 X 的无条件分布.

解 因为 (注意 gamma 分布的参数 θ 用倒数代替)

$$\begin{aligned}f_X(x)&=\frac{\theta^\alpha}{\Gamma(\alpha)}\int_0^\infty \lambda\mathrm{e}^{-\lambda x}\lambda^{\alpha-1}\mathrm{e}^{-\theta\lambda}\mathrm{d}\lambda=\frac{\theta^\alpha}{\Gamma(\alpha)}\int_0^\infty \lambda^\alpha\mathrm{e}^{-\lambda(x+\theta)}\mathrm{d}\lambda\\&=\frac{\theta^\alpha}{\Gamma(\alpha)}\frac{\Gamma(\alpha+1)}{(x+\theta)^{\alpha+1}}=\frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}}.\end{aligned}$$

故这是一个 Pareto 分布.

下例是第 16 章中的一个实例.

例 4.30 假设给定 $\Theta=\theta$, X 服从均值为 θ、方差为 v 的正态分布, 因此

$$f_{X|\Theta}(x|\theta)=\frac{1}{\sqrt{2\pi v}}\exp\left[-\frac{1}{2v}(x-\theta)^2\right],\quad -\infty<x<\infty,$$

而且 Θ 自身服从均值为 μ、方差为 a 的正态分布, 也就是

$$f_\Theta(\theta)=\frac{1}{\sqrt{2\pi a}}\exp\left[-\frac{1}{2a}(\theta-\mu)^2\right],\quad -\infty<\theta<\infty.$$

试确定 X 的边缘概率密度函数.

解 X 的边缘概率密度函数为

$$\begin{aligned}f_X(x)&=\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi v}}\exp\left[-\frac{1}{2v}(x-\theta)^2\right]\frac{1}{\sqrt{2\pi a}}\exp\left[-\frac{1}{2a}(\theta-\mu)^2\right]\mathrm{d}\theta\\&=\frac{1}{2\pi\sqrt{va}}\int_{-\infty}^\infty\exp\left[-\frac{1}{2v}(x-\theta)^2-\frac{1}{2a}(\theta-\mu)^2\right]\mathrm{d}\theta.\end{aligned}$$

将下面这个 θ 的配方代数恒等式的证明留给读者作为练习

$$\frac{(x-\theta)^2}{v}+\frac{(\theta-\mu)^2}{a}=\frac{a+v}{va}\left(\theta-\frac{ax+v\mu}{a+v}\right)^2+\frac{(x-\mu)^2}{a+v}.$$

因此有

$$f_X(x)=\frac{\exp\left[-\dfrac{(x-\mu)^2}{2(a+v)}\right]}{\sqrt{2\pi}(a+v)}\int_{-\infty}^{\infty}\sqrt{\frac{a+v}{2\pi va}}\exp\left[-\frac{a+v}{2va}\left(\theta-\frac{ax+v\mu}{a+v}\right)^2\right]\mathrm{d}\theta.$$

看出被积函数 (作为 θ 的函数) 是均值为 $(ax+v\mu)/(a+v)$、方差为 $(va)/(a+v)$ 的正态分布的概率密度函数. 因此积分为 1, 进而有

$$f_X(x)=\frac{\exp\left[-\dfrac{(x-\mu)^2}{2(a+v)}\right]}{\sqrt{2\pi(a+v)}},\quad -\infty<x<\infty;$$

即 X 服从均值为 μ、方差为 $a+v$ 的正态分布. □

下面的例子源自 Hayne[50], 说明如何构造这种混合分布. 特别地, 通常用连续混合分布来构造参数不确定的模型, 即参数的准确值并不知道, 但参数的取值满足某个给定的概率密度函数.

例 4.31 *在机动车辆保单进行评估时, 重要的一点是驾驶员之间的驾驶距离会各有不同. 同样, 对于同一个司机, 各年的驾驶里程也会不相同. 假设, 某个随机选取的司机的驾驶里程服从逆 Weibull 分布, 尺度参数各年的变化服从变换 gamma 分布, τ 值相同. 试对某随机选择的司机确定其在某一年的里程的分布.*

解 利用附录 A 中的参数表示, 一年的里程为逆 Weibull 分布, 参数为 Λ(代替 Θ) 和 τ, 同时尺度参数 Λ 的变换 gamma 分布的参数为 τ、θ 和 α. 边缘密度为

$$\begin{aligned}
f(x)&=\int_0^{\infty}\frac{\tau\lambda^{\tau}}{x^{\tau+1}}\mathrm{e}^{-(\lambda/x)^{\tau}}\frac{\tau\lambda^{\tau\alpha-1}}{\theta^{\tau\alpha}\Gamma(\alpha)}\mathrm{e}^{-(\lambda/\theta)^{\tau}}\mathrm{d}\lambda\\
&=\frac{\tau^2}{\theta^{\tau\alpha}\Gamma(\alpha)x^{\tau+1}}\int_0^{\infty}\lambda^{\tau+\tau\alpha-1}\exp\left[-\lambda^{\tau}(x^{-\tau}+\theta^{-\tau})\right]\mathrm{d}\lambda\\
&=\frac{\tau^2}{\theta^{\tau\alpha}\Gamma(\alpha)x^{\tau+1}}\int_0^{\infty}[y^{1/\tau}(x^{-\tau}+\theta^{-\tau})^{-1/\tau}]^{\tau+\tau\alpha-1}\mathrm{e}^{-y}\\
&\quad\times y^{\tau^{-1}-1}\tau^{-1}(x^{-\tau}+\theta^{-\tau})^{-1/\tau}\mathrm{d}y\\
&=\frac{\tau}{\theta^{\tau\alpha}\Gamma(\alpha)x^{\tau+1}(x^{-\tau}+\theta^{-\tau})^{\alpha+1}}\int_0^{\infty}y^{\alpha}\mathrm{e}^{-y}\mathrm{d}y\\
&=\frac{\tau\Gamma(\alpha+1)}{\theta^{\tau\alpha}\Gamma(\alpha)x^{\tau+1}(x^{-\tau}+\theta^{-\tau})^{\alpha+1}}=\frac{\tau\alpha\theta^{\tau}x^{\tau\alpha-1}}{(x^{\tau}+\theta^{\tau})^{\alpha+1}}.
\end{aligned}$$

第三行由代换 $y=\lambda^{\tau}(x^{-\tau}+\theta^{-\tau})$ 得到. 最后一行利用 $\Gamma(\alpha+1)=\alpha\Gamma(\alpha)$. 结果为逆 Burr 分布. 值得注意的是这种分布适用于特定的一类驾驶员, 其他驾驶员可能会有不同的 Weibull 分布的形状参数 τ, 尺度参数 Θ 的分布也不同, 因此得到不同的均值. □

4.4.6 含瑕点的风险率模型

混合分布中一个很重要的类型是含瑕点的风险率模型. 尽管分析这种混合类型的原始动机来源于生存分析中的寿命分布, 但其结果在数学上的简便性使得人们以这种方法作为混合构造新分布的有效途径.

首先引入一个随机风险变量 $\Lambda > 0$, 并定义 X 的条件损失率 (给定 $\Lambda = \lambda$) 为 $h_{X|\Lambda}(x\,|\lambda) = \lambda a(x)$, $a(x)$ 是 x 的一个已知函数 (即 $a(x)$ 被指定为一种特定的应用). 这里所说的含瑕点是指用一个随机风险变量对损失率函数本身的不确定性进行量化, 即表现为上面的乘积条件损失率.

$X|\Lambda$ 的条件生存函数为

$$S_{X|\Lambda}(x|\lambda) = \mathrm{e}^{-\int_0^x h_{X|\Lambda}(t|\lambda)\mathrm{d}t} = \mathrm{e}^{-\lambda A(x)},$$

其中 $A(x) = \int_0^x a(t)\mathrm{d}t$. 为了确定混合分布 (即 X 的边缘分布), 我们定义随机风险变量 Λ 的矩生成函数为 $M_\Lambda(t) = \mathrm{E}(\mathrm{e}^{t\Lambda})$. 则边缘生存函数为

$$S_X(x) = \mathrm{E}[\mathrm{e}^{-\Lambda A(x)}] = M_\Lambda[-A(x)], \tag{4.5}$$

显然有 $F_X(x) = 1 - S_X(x)$.

各种类型的混合将决定 $a(x)$ 和 $A(x)$ 的选择. 含瑕点的风险率模型的一个重要子类是混合指数分布: $a(x) = 1$ 和 $A(x) = x$, 因此 $S_{X|\Lambda}(x\,|\lambda) = \mathrm{e}^{-\lambda x}$, $x \geqslant 0$. 其他常用的混合包括混合 Weibull 分布: $a(x) = \gamma x^{\gamma-1}$ 和 $A(x) = x^\gamma$.

估计一个含瑕点的风险率分布需要已知 Λ 的矩生成函数 $M_\Lambda(t)$ 的表达式. 最常用的选择是 gamma 含瑕点风险率模型, 但其他选择如逆高斯含瑕点风险率模型也很常用.

例 4.32 *设 Λ 服从 gamma 分布, $X|\Lambda$ 服从 Weibull 分布, 其条件生存函数为 $S_{X|\Lambda}(x\,|\lambda) = \mathrm{e}^{-\lambda x^\gamma}$. 确定 X 的无条件分布或边缘分布.*

解 由条件及例 3.15 得知 gamma 分布的矩生成函数为 $M_\Lambda(t) = (1-\theta t)^{-\alpha}$, 并由 (4.5) 式得到 X 的生存函数为

$$S_X(x) = M_\Lambda(-x^\gamma) = (1+\theta x^\gamma)^{-\alpha}.$$

这是一个 Burr 分布 (见附录 A), 其参数 θ 被 $\theta^{-1/\gamma}$ 所替换. 注意到当 $\gamma = 1$ 时为指数混合分布即 Pareto 分布, 参考前面的例 4.29. □

如前所述, 通过混合可以构造尾部较厚的分布. 特别地, 对风险率递减的分布进行混合得到的分布的风险率也递减. 习题 4.32 要求读者证明含瑕点的风险率变量具有这一性质. 含瑕点的风险率模型的进一步研究可以参见 Hougaard 的著作 [63].

4.4.7 分段

构造新分布的另一种方法是分段方法, 与混合处理有些相似, 都是由两个或多个已知分布产生总损失的分布. 在混合处理中, 每个分布针对总体的一个子集. 当某个子集确认后, 就可以采用简单损失模型方法. 在分段处理中将根据损失量的不同采用不同的分布, 即一个模型只在某个损失区间上发生作用, 其他模型对应其他的区间. 更精确的定义如下.

定义 4.33 *k 元分段分布*(k-component spliced distribution) *的密度函数可以表示如下*[①]:

$$f_X(x)=\begin{cases} a_1 f_1(x), & c_0<x<c_1,\\ a_2 f_2(x), & c_1<x<c_2,\\ \quad\vdots & \quad\vdots\\ a_k f_k(x), & c_{k-1}<x<c_k. \end{cases}$$

对于每个 $a_j>0$, $j=1,\cdots,k$, $f_j(x)$ 为区间 (c_{j-1},c_j) 上全概率的合理密度函数. 并且有 $a_1+\cdots+a_k=1$.

例 4.34 *证明 2.2 节的模型 5 是一个二元分段分布.*

解 其密度函数为

$$f(x)=\begin{cases} 0.01, & 0\leqslant x<50,\\ 0.02, & 50\leqslant x<75. \end{cases}$$

这个分段模型是这样构造的: 令 $f_1(x)=0.02$, $0\leqslant x<50$, 即区间 [0,50) 上的均匀分布; $f_2(x)=0.04$, $50\leqslant x<75$, 即区间 [50,75) 上的均匀分布. 系数为 $a_1=a_2=0.5$. □

虽然在这种构造中不必考虑整个密度函数和系数, 但是这种方法可以确保最终的结果是合理的密度函数. 对于参数模型, 考虑分段的一个动机是分布的尾部厚度与小损失量时的表现很不一致. 例如, 经验 (基于现有信息, 如一小部分数据集合) 显示尾分布服从 Pareto 分布, 但同时存在一个正的众数在附近更多的像是对数正态分布或逆高斯分布. 另一个例子是, 大量的数据低于某个数值, 而对其他点的观测只有很少的信息. 我们可以在那个数值点之上采用经验分布 (或它的光滑形式), 而在这个点的下边采用参数模型. 另外, 上面定义中的间断点 $c_0,\cdots,c_k$ 应该是提前已知的.

构造分段分布的另一种方法是利用位于 c_0 到 c_k 上的标准分布. 令 $g_j(x)$ 是第 j 个这样的密度函数. 则由定义 4.33 用 $g_j(x)/[G(c_j)-G(c_{j-1})]$ 代替 $f_j(x)$, 这个公式可以使间断点成为可以估计的参数.

---

① 这个定义没有对区间端点给出定义. —— 译者注

分段处理并不保证生成的概率密度函数连续 (间断点处光滑). 当然也可以在构造中加入这种限制.

例 4.35　利用 $(0,c)$ 区间上的指数分布和 (c,∞) 区间上的 Pareto 分布 (用 γ 代替参数 θ) 构造一个二元分段分布.

解　基本公式为

$$f_X(x)=\begin{cases} a_1\dfrac{\theta^{-1}\mathrm{e}^{-x/\theta}}{1-\mathrm{e}^{-c/\theta}}, & 0<x<c,\\ a_2\dfrac{\alpha\gamma^{\alpha}(x+\gamma)^{-\alpha-1}}{\gamma^{\alpha}(c+\gamma)^{-\alpha}}, & c<x<\infty.\end{cases}$$

然而我们必须使密度函数的积分为 1, 这只要使 $a_1=v, a_2=1-v$ 即可. 则分段接合的密度函数为

$$f_X(x)=\begin{cases} v\dfrac{\theta^{-1}\mathrm{e}^{-x/\theta}}{1-\mathrm{e}^{-c/\theta}}, & 0<x<c,\\ (1-v)\dfrac{\alpha(c+\gamma)^{\alpha}}{(x+\gamma)^{\alpha+1}}, & c<x<\infty,\end{cases}\quad \theta,\alpha,\gamma,c>0,\quad 0<v<1.$$

图 4-4 显示了参数为 $c=100$, $v=0.6$, $\theta=100$, $\gamma=200$ 和 $\alpha=4$ 的密度函数, 显然这个密度函数并不连续. □

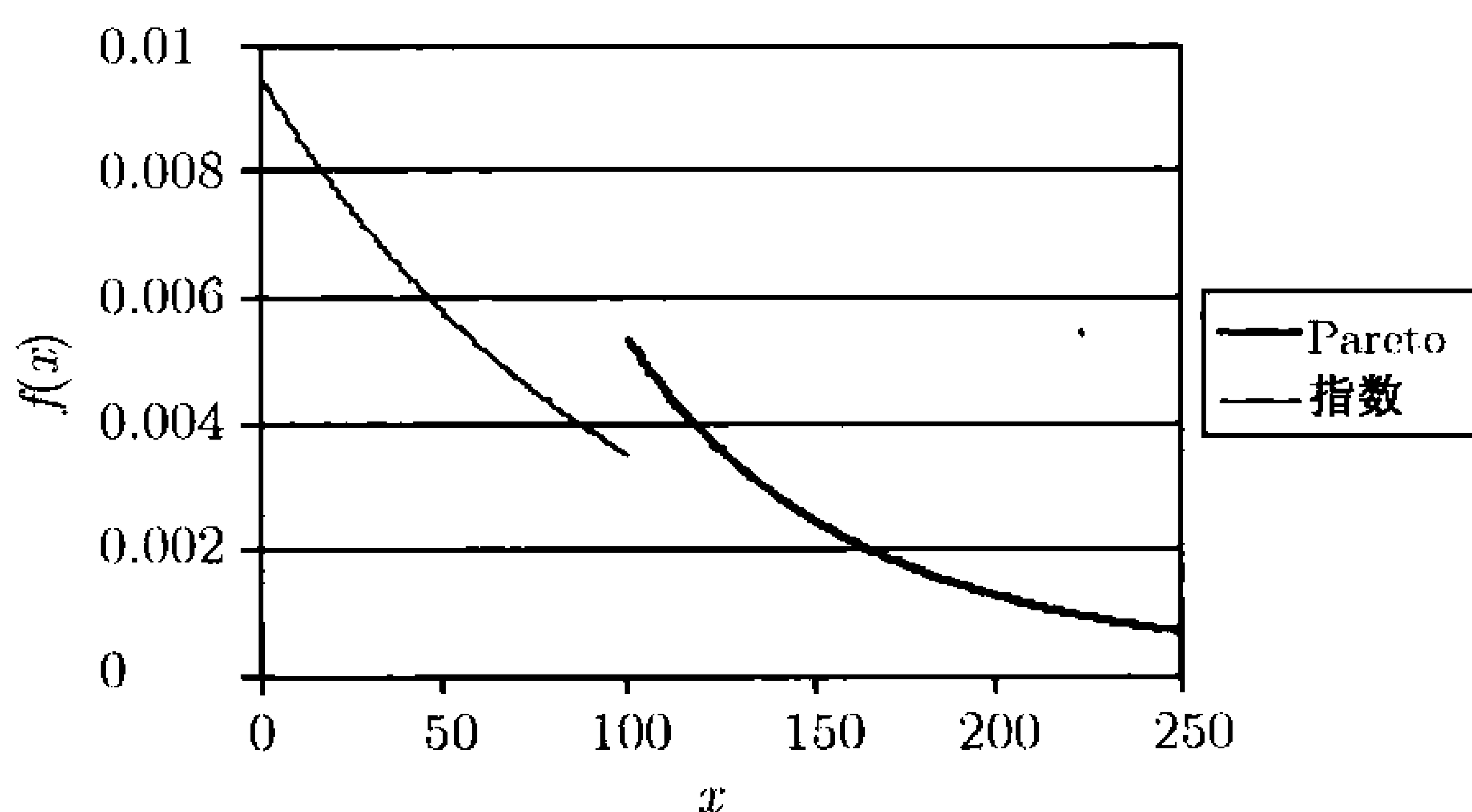


图 4-4　二元分段接合的分布密度函数

习题

4.18　令 X 的累积分布函数为 $F_X(x)=1-(1+x)^{-\alpha}, x,\alpha>0$. 确定 $Y=\theta X$ 的概率密度函数和累积分布函数.

4.19*　1995 年中的 100 个观测赔案的索赔金额如下: 0~300 有 42 个, 300~350 有 3 个, 350~400 有 5 个, 400~450 有 5 个, 450~500 没有, 500~600 有 5 个, 超过 600 有 40 个. 接下来的 3 年中, 索赔金额每年增长 10%. 基于 1995 年的经验分布, 确定 1998 年索赔金额超过 500 的概率范围 (给出的信息不足以推出准确的概率值).

4.20　X 服从 Pareto 分布. 确定其变换分布、逆分布和逆变换分布的累积分布函数. 并对照附录 A 确定它们是否有特殊的名称.

4.21 X 服从对数 logistic 分布. 证明其逆分布也服从对数 logistic 分布. 因此不需要再定义逆对数 logistic 分布.

4.22 Y 服从参数为 μ 和 σ 的对数正态分布. 令 $Z = \theta Y$. 证明 Z 服从对数正态分布, 因此引入第三个参数并没有产生新的分布.

4.23* X 服从参数是 α 和 θ 的 Pareto 分布. 令 $Y = \ln(1 + X/\theta)$. 确定 Y 的分布类型和参数.

4.24 在文献 [132] 中 Venter 指出若 X 服从变换 gamma 分布, 其尺度参数 θ 服从逆变换 gamma 分布 (两个分布的参数都为 τ), 则混合的结果是变换 beta 分布. 证明这一结论.

4.25* N 服从均值为 Λ 的 Poisson 分布. Λ 服从均值为 1 方差为 2 的 gamma 分布. 计算 $N = 1$ 的无条件概率.

4.26* 给定 $\Theta = \theta$, 随机变量 X 服从指数分布, 其损失率函数为 $h(x) = \theta$(常数). 随机变量 Θ 服从 (1,11) 上的均匀分布. 计算无条件分布的函数值 $S_X(0.5)$.

4.27* N 服从均值为 Λ 的 Poisson 分布. Λ 服从 (0,5) 上的均匀分布. 计算 $N \geqslant 2$ 的无条件概率.

4.28 确定含瑕点的风险率分布的概率密度函数和损失率函数.

4.29 假设 $X|\Lambda$ 服从 Weibull 生存函数 $S_{X|\Lambda}(x|\lambda) = \mathrm{e}^{-\lambda x^{\gamma}}, x \geqslant 0$, Λ 服从指数分布. 证明 X 的无条件分布为对数 logistic 分布.

4.30 考虑指数逆高斯随机风险率模型, 满足 $a(x) = \theta/(2\sqrt{1+\theta x})$, $\theta > 0$.

(a) 证明 $X|\Lambda$ 的条件损失率函数 $h_{X|\Lambda}(x|\lambda)$ 的确为一个有效的损失率函数.

(b) 确定条件生存函数 $S_{X|\Lambda}(x|\lambda)$.

(c) 若 Λ 服从 gamma 分布, 参数为 $\theta = 1$, α 由 2α 代替. 确定 X 的边缘或无条件生存函数.

(d) 利用 (c) 证明一个给定的含瑕点风险率模型可以由多个 $X|\Lambda$ 的条件分布和随机风险分布 Λ 结合产生.

4.31 假设 X 的生存函数为 $S_X(x) = 1 - F_X(x)$, 由 (4.5) 式给出. 证明 $S_1(x) = F_X(x)/[\mathrm{E}(\Lambda)A(x)]$ 是 (4.5) 的另一种生存函数, 并利用 $S_1(x)$ 给出 Λ 的分布.

4.32 给定 $s \geqslant 0$, 定义一个 "Esscher 变换" 含瑕点的风险率随机变量 Λ_s, 其概率密度函数 (或是离散分布概率) 为 $f_{\Lambda_s}(\lambda) = \mathrm{e}^{-s\lambda} f_\Lambda(\lambda)/M_\Lambda(-s), \lambda \geqslant 0$.

(a) 证明 Λ_s 的矩生成函数为

$$M_{\Lambda_s}(t) = \mathrm{E}(\mathrm{e}^{t\Lambda_s}) = \frac{M_\Lambda(t-s)}{M_\Lambda(-s)}.$$

(b) 定义 Λ 的累积生成函数为

$$c_\Lambda(t) = \ln[M_\Lambda(t)],$$

利用 (a) 证明

$$c'_\Lambda(-s) = \mathrm{E}(\Lambda_s), \quad c''_\Lambda(-s) = \mathrm{Var}(\Lambda_s).$$

(c) 对于含瑕点的风险率模型, 其生存函数由 (4.5) 式给出, 证明其相关的损失率为 $h_X(x)=a(x)c'_\Lambda[-A(x)]$, c_Λ 在 (b) 中定义.

(d) 利用 (c) 证明

$$h'_X(x)=a'(x)c'_\Lambda[-A(x)]-[a(x)]^2c''_\Lambda[-A(x)].$$

(e) 利用 (d) 证明, 若条件损失率 $h_{X|\Lambda}(x|\lambda)$ 关于 x 非增, 则 $h_X(x)$ 也关于 x 非增.

4.33 写出一个二元分段接合模型的密度函数. 其密度函数在 (0, 1 000) 上正比于均匀分布, 在 (1 000, ∞) 上正比于指数分布密度函数. 并且要确保生成的密度函数连续.

4.34 X 的概率密度函数为 $f(x)=\exp(-|x/\theta|)/2\theta, -\infty<x<\infty$. 令 $Y=\mathrm{e}^X$, 确定 Y 的概率密度函数和累积分布函数.

4.35* 1993 年的损失服从密度函数 $f(x)=3x^{-4}, x\geqslant 1$, 其中 x 是以百万美元为单位计量的损失量. 由于通货膨胀, 损失量的增长率为 10%, 从 1993 年到 1994 年均匀增长. 确定 1994 年损失量的累积分布函数, 并用其确定 1994 年损失超过 2 200 000 的概率.

4.36 考虑一个逆高斯分布随机变量 X, 其概率密度函数 (由附录 A) 为

$$f(x)=\sqrt{\frac{\theta}{2\pi x^3}}\exp\left[-\frac{\theta}{2x}\left(\frac{x-\mu}{\mu}\right)^2\right],\quad x>0,$$

参数 $\theta>0$, $\mu>0$.

(a) 推导出这个逆高斯随机变量的倒数 $1/X$ 的概率密度函数.

(b) 证明 X 和 $1/X$ 的联合矩生成函数为

$$\begin{aligned}M(t_1,t_2)&=\mathrm{E}\left(\mathrm{e}^{t_1X+t_2X^{-1}}\right)\\&=\sqrt{\frac{\theta}{\theta-2t_2}}\exp\left(\frac{\theta-\sqrt{(\theta-2\mu^2t_1)(\theta-2t_2)}}{\mu}\right),\end{aligned}$$

其中 $t_1<\theta/(2\mu^2), t_2<\theta/2$

(c) 利用 (b) 证明 X 的矩生成函数为

$$M_X(t)=\mathrm{E}(\mathrm{e}^{tX})=\exp\left[\frac{\theta}{\mu}\left(1-\sqrt{1-\frac{2\mu^2}{\theta}t}\right)\right],\quad t<\frac{\theta}{2\mu^2},$$

这与习题 3.24 的重新参数化的结果是一致的.

(d) 利用 (b) 证明逆高斯随机变量的倒数 $1/X$ 的矩生成函数为

$$M_{1/X}(t)=\mathrm{E}\left(\mathrm{e}^{tX^{-1}}\right)=\sqrt{\frac{\theta}{\theta-2t}}\exp\left[\frac{\theta}{\mu}\left(1-\sqrt{1-\frac{2}{\theta}t}\right)\right],\quad t<\frac{\theta}{2}.$$

由此证明 $1/X$ 与 Z_1+Z_2 同分布, 其中 Z_1 服从 gamma 分布, Z_2 服从逆高斯分布, 且二者独立. 在这种形式下给出两个分布的参数.

(e) 利用 (b) 证明

$$Z = \frac{1}{X}\left(\frac{X-\mu}{\mu}\right)^2$$

服从 gamma 分布, 参数为 $\alpha = \frac{1}{2}$, 用 $2/\theta$ 替换 θ(如附录 A).

4.5 常用分布及其相互关系

4.5.1 引言

有很多种方法可以对分布归类. 分布族的讨论见参考文献 [73] 的第 2 章, 如 Pearson(12 种), Burr(12 种), Stoppa(5 种), Dagum(11 种). 同一个分布可能会出现在不止一个体系中, 这说明分布之间存在复杂的关系, 不只是形式上所表达的那样. 4.5.2 节的介绍对于精算建模十分有用, 因为所有分布的支集都位于正实数集, 并向右倾斜. 对于由所有连续分布构成的集合, John, Kots 和 Balakrishnan 的两部著作 ([67] 和 [68]) 是很好的参考书. 另外还有专门介绍个别分布的书籍 (如 Arnold 的著作 [5] 介绍的 Pareto 分布).

4.5.2 两参数分布族

在定义参数分布族时, 本节列出的很多分布 (包括附录 A) 都是某些更一般分布的特例. 例如, $\tau = 1, \theta$ 任意的 Weibull 分布是指数分布. 通过这种方法, 很多分布可以被归为一类, 如图 4-5 和图 4-6. 变换 beta 分布族包含两种性质完全不同的特殊情形, Paralogistic 分布和逆 Paralogistic 分布只需在 Burr 分布和逆 Burr 分布中各令两个非尺度参数相等, 而不一定为某个给定的值.

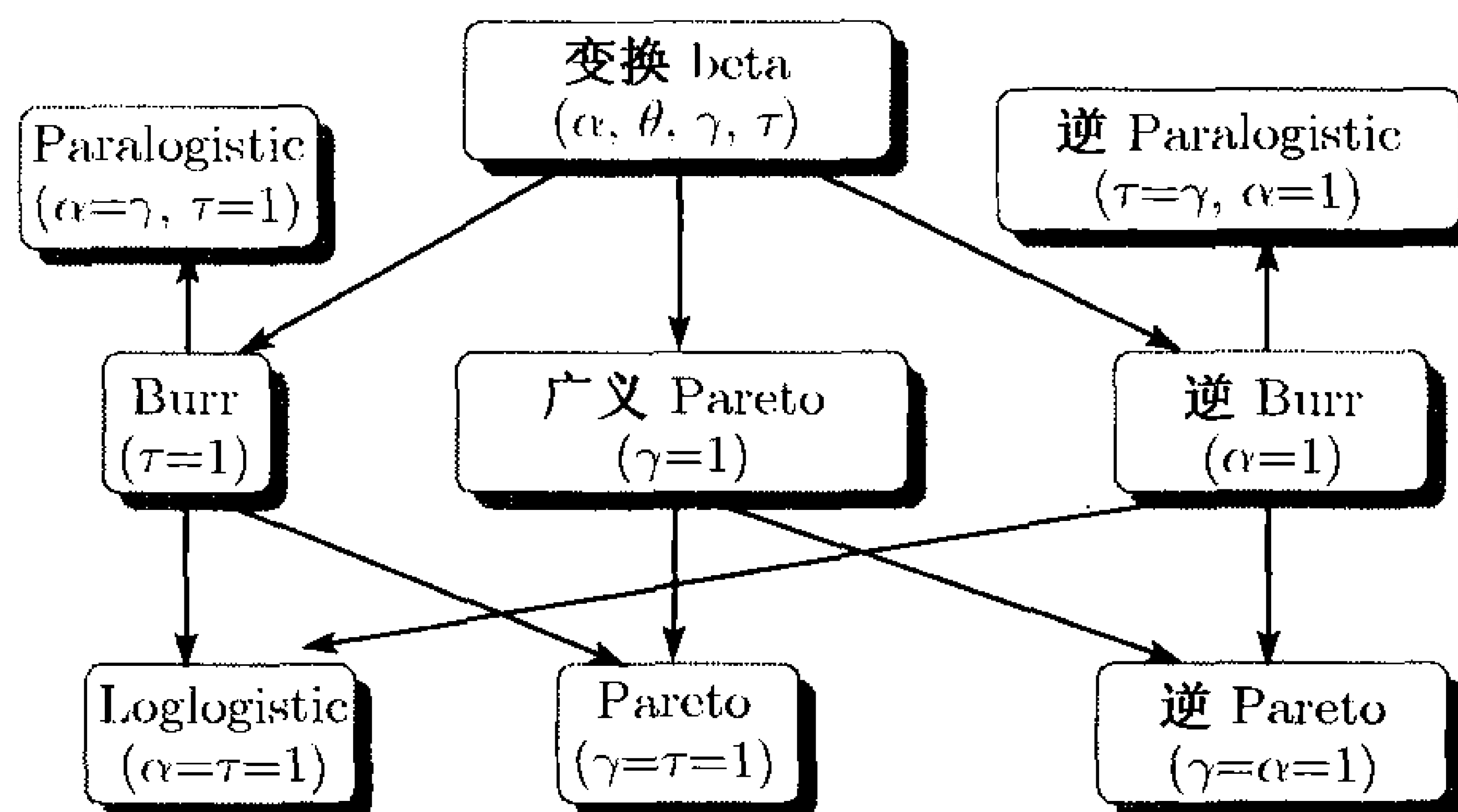


图 4-5 变换 beta 分布族

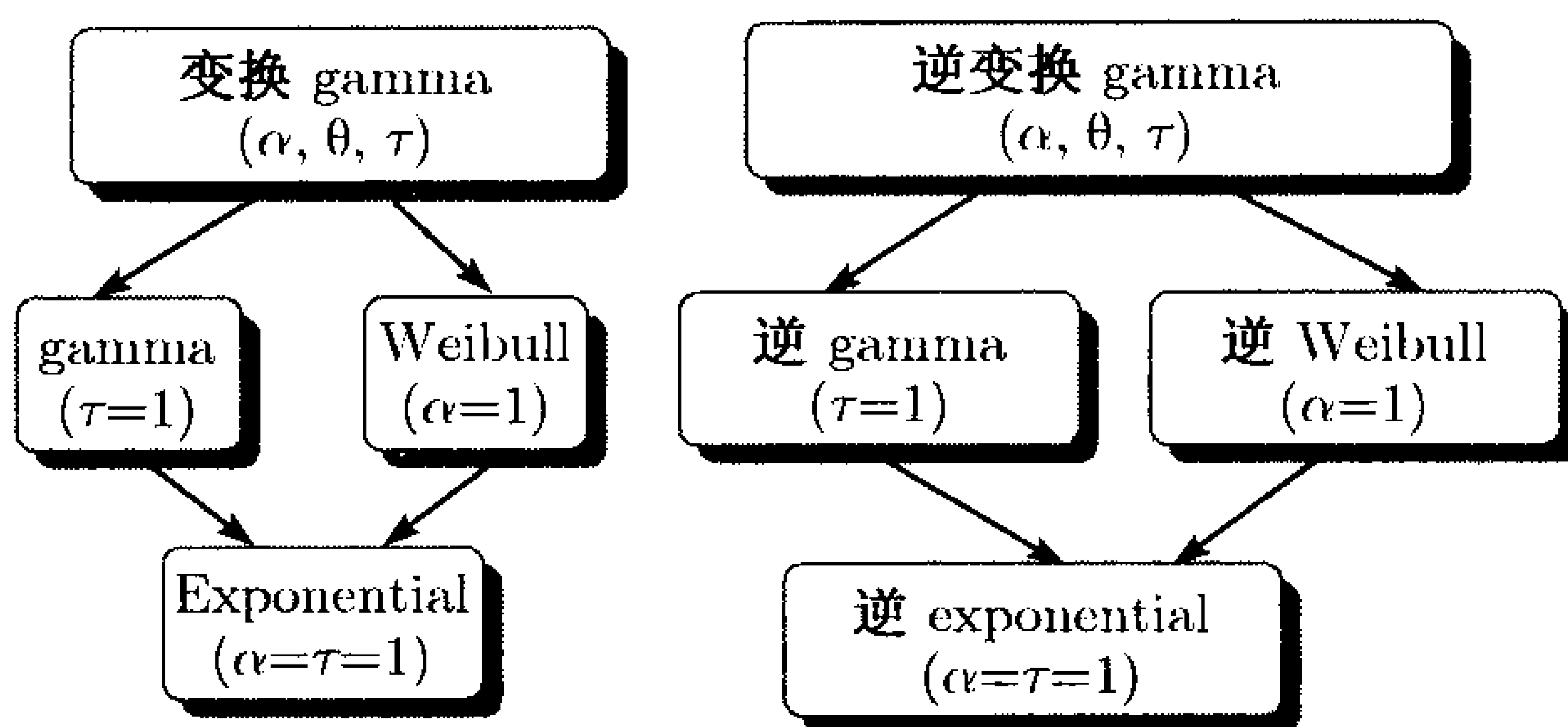


图 4-6 变换/逆变换 gamma 分布族

4.5.3 分布的极限

4.5.2 节讨论的分类方法可能会使得某些分布是其他分布的特例. 另一种研究各类分布关系的方法是考察参数趋于零或无穷大时分布的极限情况.

例 4.36 证明变换 gamma 分布是变换 beta 分布当 $\theta\to\infty$, $\alpha\to\infty$ 和 $\theta/\alpha^{1/\gamma}\to\xi$(某个常数) 时的极限分布.

证明 证明依赖以下两个关于极限的结论

$$\lim_{\alpha\to\infty}\frac{\mathrm{e}^{-\alpha}\alpha^{\alpha-1/2}(2\pi)^{1/2}}{\Gamma(\alpha)}=1, \tag{4.6}$$

$$\lim_{a\to\infty}\left(1+\frac{x}{a}\right)^{a+b}=\mathrm{e}^{x}. \tag{4.7}$$

(4.6) 式中的极限为著名的 Stirling 公式给出 gamma 函数的某种近似. (4.7) 式中的极限为一般微积分教材中的一个标准结果.

为确保比率 $\theta/\alpha^{1/\gamma}$ 趋向于某一常数, 只需在 α 和 θ 不断增大时固定其比例. 在变换 beta 分布的概率密度函数中, 用 $\xi\alpha^{1/\gamma}$ 代替 θ, 并令 $\alpha\to\infty$. 第一步, 用 Stirling 公式替换两个 gamma 函数的形式

$$\begin{aligned}
f(x)&=\frac{\Gamma(\alpha+\tau)\gamma x^{\gamma\tau-1}}{\Gamma(\alpha)\Gamma(\tau)\theta^{\gamma\tau}(1+x^{\gamma}\theta^{-\gamma})^{\alpha+\tau}}\\
&=\frac{\mathrm{e}^{-\alpha-\tau}(\alpha+\tau)^{\alpha+\tau-1/2}(2\pi)^{1/2}\gamma x^{\gamma\tau-1}}{\mathrm{e}^{-\alpha}\alpha^{\alpha-1/2}(2\pi)^{1/2}\Gamma(\tau)(\xi\alpha^{1/\gamma})^{\gamma\tau}(1+x^{\gamma}\xi^{-\gamma}\alpha^{-1})^{\alpha+\tau}}\\
&=\frac{\mathrm{e}^{-\tau}[(\alpha+\tau)/\alpha]^{\alpha+\tau-1/2}\gamma x^{\gamma\tau-1}}{\Gamma(\tau)\xi^{\gamma\tau}[1+(x/\xi)^{\gamma}/\alpha]^{\alpha+\tau}}.
\end{aligned}$$

将两个极限

$$\lim_{\alpha\to\infty}\left(1+\frac{\tau}{\alpha}\right)^{\alpha+\tau-1/2}=\mathrm{e}^{\tau},\quad \lim_{\alpha\to\infty}\left[1+\frac{(x/\xi)^{\gamma}}{\alpha}\right]^{\alpha+\tau}=\mathrm{e}^{(x/\xi)^{\gamma}}$$

代入后得到

$$\lim_{\alpha\to\infty} f(x) = \frac{\gamma x^{\gamma\tau-1}\mathrm{e}^{-(x/\xi)^{\gamma}}}{\Gamma(\tau)\xi^{\gamma\tau}},$$

即为变换 gamma 分布的概率密度函数. □

采用类似的方法, 逆变换 gamma 分布可以令 τ 趋于无穷大而得到 (见习题 4.39).

因为 Burr 分布是 $\tau=1$ 的变换 beta 分布的特例, 它的极限情况是 $\tau=1$ 的变换 gamma 分布的特例 (使用上例中的参数化方法), 即 Weibull 分布. 类似地, 逆 Weibull 分布是逆 Burr 分布的一个极限情况. 最后, 令 $\tau=\gamma=1$, 得到指数分布是 Pareto 分布的一个极限情况 (逆分布类似).

作为极限情况的最后一个说明, 考虑上面提过的参数化的变换 gamma 分布. 令 $\gamma^{-1}\sqrt{\xi^{\gamma}}\to\sigma$ 和 $\gamma^{-1}(\xi^{\gamma}\tau-1)\to\mu$, 再令 τ 趋于无穷大 (所以 γ,ξ 一定都趋向于 0) 得到的极限分布为对数正态分布.

图 4-7 中列出了一些分布的极限及其特殊关系的情况. 也给出了关于不同分布间其他一些有趣的事实[①].

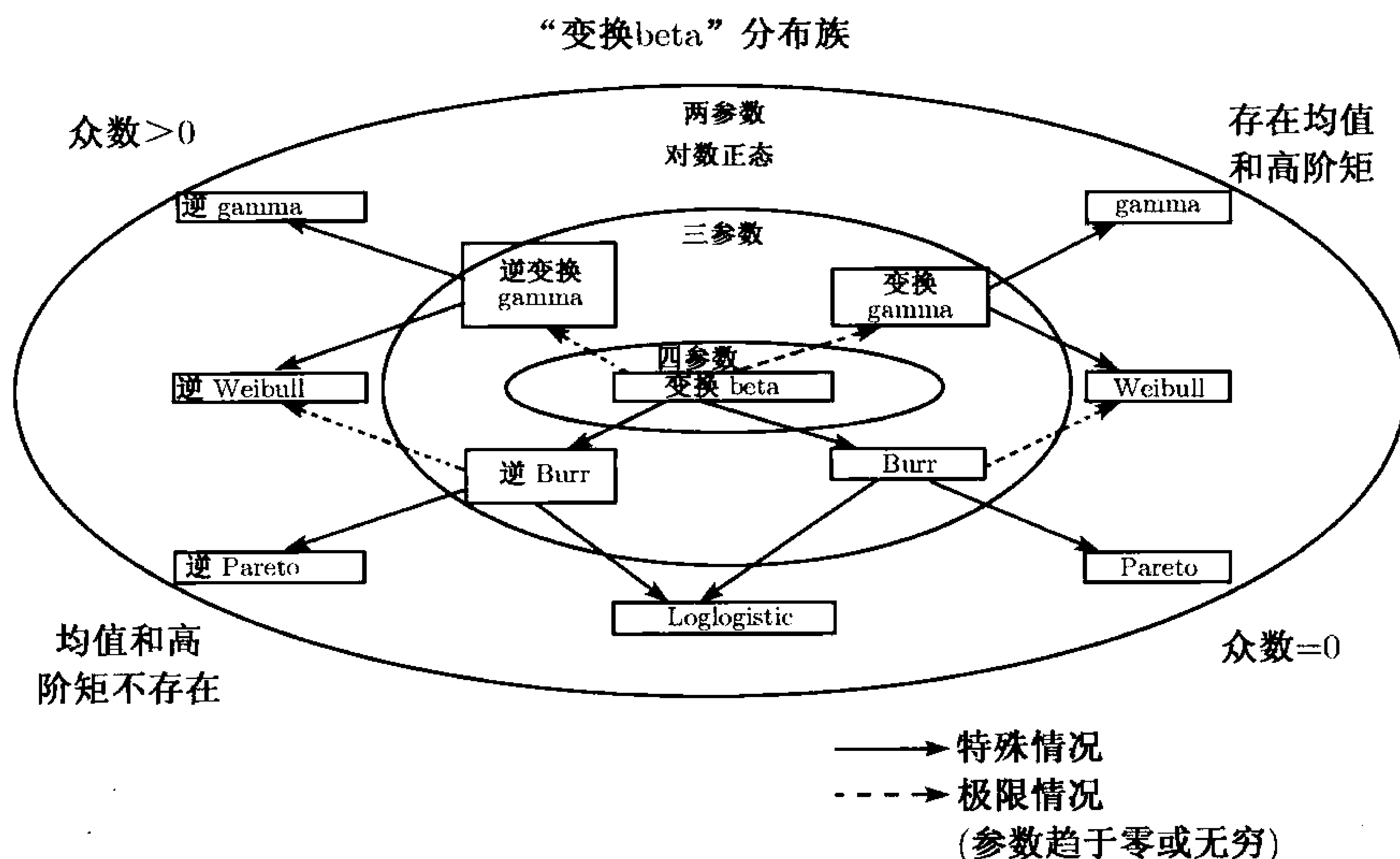


图 4-7 分布的关系和特点

习题

4.37 对于 Pareto 分布, 保持 α/θ 为常数, 证明当 α 和 θ 趋于无穷大时极限为指数分布.

4.38 确定广义 Pareto 分布当 α 和 θ 趋于无穷大时的极限分布.

4.39 证明当 $\tau\to\infty$ 时的变换 beta 分布为逆变换 gamma 分布.

① 感谢美国 Re-Insurance 公司的 Dave Clark 制作此图.

4.6　离散分布

4.6.1　引言

本节将讨论计数分布这一大类分布. 计数分布为仅在非负整数上有概率的离散分布, 即概率仅定义在点 0,1,2,3,4, $\cdots$ 上. 在保险的背景下, 计数分布用来描述事件的索赔次数, 这里的事件可以是保险损失的发生也可以是保险索赔的发生. 相比于仅知道总损失量, 对索赔次数和规模都有所了解, 可以更加深入地理解与承保相关的各种因素. 将总损失量分别用理赔次数和索赔额表示将有助于我们对保单契约的修改. 进行这样表示的另一个原因是索赔次数的模型比较容易获得, 并且经验表明常用的分布也的确分别适用于不同的损失模型.

现在对离散模型中要用到的一些符号形式化. **概率函数**(probability function, pf)p_k 表示恰巧发生 k 次事件 (索赔或损失) 的概率. 令 N 表示这类事件发生次数的随机变量. 则

$$p_k = \Pr(N = k), \quad k = 0, 1, 2, \cdots .$$

注意, 离散随机变量 N 的概率生成函数 (pgf) 可以用 p_k 表示为

$$P(z) = P_N(z) = \mathrm{E}(z^N) = \sum_{k=0}^{\infty} p_k z^k. \tag{4.8}$$

与矩生成函数类似, 也可以由概率生成函数得到随机变量的矩. 特别地, 有 $P'(1) = \mathrm{E}(N), P''(1) = \mathrm{E}[N(N-1)]$(习题 4.42). 并且可以由概率生成函数得到各个概率

$$\begin{aligned} P^{(m)}(z) &= \mathrm{E}\left(\frac{\mathrm{d}^m}{\mathrm{d}z^m} z^N\right) = \mathrm{E}[N(N-1)\cdots(N-m+1)z^{N-m}] \\ &= \sum_{k=m}^{\infty} k(k-1)\cdots(k-m+1) z^{k-m} p_k, \\ P^{(m)}(0) &= m! p_m \text{ 或 } p_m = \frac{P^{(m)}0}{m!}. \end{aligned}$$

4.6.2　Poisson 分布

Poisson 分布的概率函数为

$$p_k = \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, 2, \cdots .$$

由例 3.16 得到的概率生成函数为

$$P(z) = \mathrm{e}^{\lambda(z-1)}, \quad \lambda > 0.$$

由概率生成函数可以得到均值和方差如下：

$$\begin{aligned}
\mathrm{E}(N) &= P'(1) = \lambda, \\
\mathrm{E}[N(N-1)] &= P''(1) = \lambda^2, \\
\mathrm{Var}(N) &= \mathrm{E}[N(N-1)] + \mathrm{E}(N) - [\mathrm{E}(N)]^2 \\
&= \lambda^2 + \lambda - \lambda^2 \\
&= \lambda.
\end{aligned}$$

Poisson 分布的均值和方差相等. Poisson 分布可以由 Poisson 过程产生 (见第 8 章的讨论). Poisson 分布和 Poisson 过程在很多统计和精算学的教科书中都有研究, 包括 Panjer and Willmot[106] 以及 Ross[116][①].

Poisson 分布至少有两个有价值的性质, 第一个是下面的定理.

定理 4.37 令 $N_1, \cdots, N_n$ 分别是参数为 $\lambda_1, \cdots, \lambda_n$ 的独立 Poisson 变量. 则 $N = N_1 + \cdots + N_n$ 服从参数为 $\lambda = \lambda_1 + \cdots + \lambda_n$ 的 Poisson 分布.

证明 独立随机变量和的概率生成函数是个体概率生成函数的乘积. 则对于 Poisson 变量的和, 我们有

$$P_N(z) = \prod_{j=1}^{n} P_{N_j}(z) = \prod_{j=1}^{n} \exp[\lambda_j(z-1)] = \exp\left[\sum_{j=1}^{n} \lambda_j(z-1)\right] = \mathrm{e}^{\lambda(z-1)},$$

其中 $\lambda = \lambda_1 + \cdots + \lambda_n$. 与矩生成函数一样, 每个分布的概率生成函数是唯一的, 因此 N 服从参数是 λ 的 Poisson 分布. □

第二个性质对保险风险建模尤为有用. 假设索赔的数量在一定时间内, 比如一年, 服从 Poisson 分布. 进一步假设索赔可以分成 m 种不同的类型. 比如, 索赔可以根据规模分类, 低于某个值为一类, 超过这个值为另一类. 可以证明, 超过某个特定值的索赔个数也服从 Poisson 分布, 但 Poisson 参数将改变.

当设计保险产品时考虑增加或去掉一部分承保责任时, 这个性质将十分有用. 假设已知某复杂的医疗保险合同的索赔次数服从 Poisson 分布, 现在考虑将索赔按照治疗方法或实际赔付金额进行 “分类”, 这时如果将某个索赔从该计划中移除后, 可以证明修改后的计划的索赔次数依然服从 Poisson 分布, 只是 Poisson 参数发生了变化.

在前面讨论的结论中, 不同类型责任的索赔次数不仅是服从 Poisson 分布而且还是相互独立的, 即低于某个索赔额的索赔次数与超过那个索赔额的索赔次数的发生是独立的, 这是一个令人有些惊讶的结果. 例如, 假设目前在销售免赔额为 50 的

① 中文版和英文影印版《应用随机过程：概率模型导论 (第 9 版)》已由人民邮电出版社出版.
—— 编者注

保单, 经验表明赔付次数服从 Poisson 分布. 进一步假设在某个阶段内的损失次数服从 Poisson 分布, 但参数未知, 若没有其他额外的信息, 不可能得到当免赔额降低或完全去掉后索赔次数的 Poisson 参数. 我们将这些想法归为如下定理.

定理 4.38 假设事件发生次数 N 服从参数为 λ 的 Poisson 分布. 进一步假设每个事件都属于 m 种类型之一, 概率分别为 $p_1,\cdots,p_m$ 且相互独立. 则属于类型 $1,2,\cdots,m$ 的事件个数 $N_1,\cdots,N_m$ 相互独立且服从均值为 $\lambda p_1,\cdots,\lambda p_m$ 的 Poisson 分布.

证明 固定 $N=n$, 条件联合分布 $(N_1,\cdots,N_m)$ 为参数 $(n,p_1,\cdots,p_m)$ 的多项式分布. 同样, 固定 $N=n$, N_j 条件边缘分布为参数 (n,p_j) 的二项式分布, $j=1,2,\cdots,m$.

$(N_1,\cdots,N_m)$ 的联合概率函数由下式给出

$$\begin{aligned}\Pr(N_1=n_1,\cdots,N_m=n_m)&=\Pr(N_1=n_1,\cdots,N_m=n_m|N=n)\times\Pr(N=n)\\&=\frac{n!}{n_1!n_2!\cdots n_m!}p_1^{n_1}\cdots p_m^{n_m}\frac{\mathrm{e}^{-\lambda}\lambda^n}{n!}\\&=\prod_{j=1}^m\mathrm{e}^{-\lambda p_j}\frac{(\lambda p_j)^{n_j}}{n_j!},\end{aligned}$$

其中 $n=n_1+n_2+\cdots+n_m$. 类似地, N_j 的边缘概率函数由下式确定

$$\begin{aligned}\Pr(N_j=n_j)&=\sum_{n=n_j}^{\infty}\Pr(N_j=n_j|N=n)\Pr(N=n)\\&=\sum_{n=n_j}^{\infty}\begin{pmatrix}n\\n_j\end{pmatrix}p_j^{n_j}(1-p_j)^{n-n_j}\frac{\mathrm{e}^{-\lambda}\lambda^n}{n!}\\&=\mathrm{e}^{-\lambda}\frac{(\lambda p_j)^{n_j}}{n_j!}\sum_{n=n_j}^{\infty}\frac{[\lambda(1-p_j)]^{n-n_j}}{(n-n_j)!}\\&=\mathrm{e}^{-\lambda}\frac{(\lambda p_j)^{n_j}}{n_j!}\mathrm{e}^{\lambda(1-p_j)}=\mathrm{e}^{-\lambda p_j}\frac{(\lambda p_j)^{nj}}{n_j!}.\end{aligned}$$

因此联合概率函数是所有边缘概率函数的乘积, 故互相独立. □

例 4.39 在医疗保险中, 索赔次数服从均值为 2.3 的 Poisson 分布. 考虑将保险责任中去掉某种治疗方法. 根据历史经验, 这种治疗方法的索赔次数约占总索赔次数的 10%. 确定新的索赔频率分布.

解 由定理 4.38, 我们知道将保险责任中去掉这种治疗方法后的修正保单, 索赔次数服从均值为 0.9(2.3)=2.07 的 Poisson 分布. 要得到总损失量的分布, 包括新政策的近似保费, 需要知道损失量的变化, 即损失程度的分布, 因为去掉的治疗方法的损失量分布可能与包含所有治疗方法的损失量分布不同. □

4.6.3 负二项分布

负二项分布通常用于替换 Poisson 分布. 同 Poisson 分布一样, 它也在非负整数上取值. 因为它包含两个参数, 因此相比 Poisson 分布其变化更灵活.

定义 4.40 负二项分布 (negative binomial distribution) 的概率函数由下式给出

$$\Pr(N=k)=p_k=\binom{k+r-1}{k}\left(\frac{1}{1+\beta}\right)^r\left(\frac{\beta}{1+\beta}\right)^k,$$
$$k=0,1,2,\cdots,\quad r>0,\quad \beta>0. \tag{4.9}$$

二项系数可表示为

$$\binom{x}{k}=\frac{x(x-1)\cdots(x-k+1)}{k!}.$$

其中 k 必须是整数, x 可以为任意实数. 当 $x>k-1$ 时, 上式也可以写成

$$\binom{x}{k}=\frac{\Gamma(x+1)}{\Gamma(k+1)\Gamma(x-k+1)},$$

这个公式更实用, 因为在很多电子制表中都预存了 $\ln\Gamma(x)$ 的数值, 各种程序语言和数学软件中也都有设置.

不难证明负二项分布的概率生成函数为

$$P(z)=[1-\beta(z-1)]^{-r}.$$

由此得到负二项分布的均值和方差

$$\mathrm{E}(N)=r\beta,\quad \mathrm{Var}(N)=r\beta(1+\beta).$$

因为 β 是正值, 所以负二项分布的方差大于均值. 而 Poisson 分布的方差和均值相等. 这说明当某类数据集观测到的方差大于均值时, 负二项分布要比 Poisson 分布更合适.

负二项分布是广义的 Poisson 分布, 至少体现在两个方面, 首先, 与 gamma 分布混合后的混合 Poisson 分布是负二项分布 (证明见本小节的后面), 其次, 以对数二阶分布生成的复合 Poisson 分布也是负二项分布 (见 4.6.7 节). 关于 Poisson 分布的另一种观点见第 8 章. 这里的一个假设是索赔发生的频率在一定的时期内为一常数, 如果这个频率随已有的索赔次数线性递增, 则在任一时期内的索赔次数服从负二项分布, 具体见 Insurance Risk Models[106] 中定理 3.6.1 的推导.

几何分布是负二项分布中当 $r=1$ 时的特例. 在某种意义上, 几何分布类似于连续型指数分布的离散形式. 几何分布和指数分布都具有指数衰减的概率函数, 因此都具有无记忆性质. 无记忆性可以从以下几个不同的方面进行解释. 如果生存函

数是指数分布, 则对于任意给定的年龄, 未来的期望生存时间均为常数. 如果用指数分布来描述保险索赔的规模, 则无记忆性可以解释为: **已知索赔超过某个水平 d, 则超过 d 的期望赔付额是一个常数从而不依赖于 d.** 即考虑免赔额 d 后, 每次赔付的索赔额的期望值不变, 而索赔次数的期望值减小了. 如果几何分布用来描述保险索赔的次数, 则无记忆性可以解释为: **给定索赔次数至少为 m, 超过 m 的期望赔付次数的概率分布是一个常数从而不依赖于 m.** 在连续分布中, 可以用指数分布来区分尾部较厚 (肥) 和尾部较轻 (瘦) 的**下指数**分布. 类似地, 在索赔频率分布中, 尾部衰减慢于几何分布的分布被认为尾部较厚, 而尾部衰减快于几何分布的分布被认为尾部较轻. 负二项分布在 $r < 1$ 时尾部较长 (衰减速度慢于几何分布), 在 $r > 1$ 时尾部较短.

如前面提到的, 一种构造负二项分布的方法是混合 Poisson 分布. 假设我们已知索赔分布的风险是 Poisson 分布, 风险参数 λ 已知. 现假设 λ 是随机变量 $\varLambda$ 的一次实现. 令 $\varLambda$ 的概率函数为 $u(\lambda)$, $\varLambda$ 可以是连续或者离散的变量, 其累积分布函数为 $U(\lambda)$. 将 λ 看作是一种随机变量实现的想法来源于以下几方面. 首先, 考虑风险群体中会有不同的风险参数 $\varLambda$, 现实中这是有意义的. 考虑一组同样保费的保单, 如对象是同一类机动车驾驶员. 这一承保类的范围包括年里程为 0~7500 英里, 居住在乡村而每周往返上班行驶距离小于 50 英里, 等等. 我们知道即使同一类的驾驶员风险也不是完全相同的, 尽管从保险人的角度看保费是相同的. 参数 λ 表示事故发生的期望次数. 若全体驾驶员的 λ 各有不同, 我们考虑被保个体是所有可能司机全体中选出的样本. 这意味着保险人知道总体中 λ 服从某一个分布 $u(\lambda)$, 而不知道其具体值. λ 的真实值是不可观测的, 能够观测到的是每个个体的事故发生次数. 因此这里有一个附加的不确定因素, 即参数的不确定性.

这个混合过程和 4.4.5 节讨论的关于连续分布的混合一样. 在一些文章中称其为**参数不确定性.** 在有关贝叶斯的文献中, 称 Λ 的分布为**先验分布,** 称其分布的参数为**超参数**(hyperparameter). 分布 $U(\cdot)$ 的作用在第 16 章的信度理论中十分重要. 当参数 λ 未知时, 发生 k 次索赔的概率可以写成在 $\Lambda = \lambda$ 条件下概率的期望值. 由全概率公式, 有

$$\begin{aligned}p_k &= \Pr(N = k) = \mathrm{E}[\Pr(N = k|\Lambda)]\\ &= \int_0^{\infty} \Pr(N = k|\Lambda = \lambda)u(\lambda)\mathrm{d}\lambda = \int_0^{\infty} \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}u(\lambda)\mathrm{d}\lambda.\end{aligned}$$

现在假设 Λ 服从 gamma 分布, 则

$$p_k = \int_0^{\infty} \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}\frac{\lambda^{\alpha-1}\mathrm{e}^{-\frac{\lambda}{\theta}}}{\theta^{\alpha}\Gamma(\alpha)}\mathrm{d}\lambda = \frac{1}{k!} = \frac{1}{\theta^{\alpha}\Gamma(\alpha)}\int_0^{\infty} \mathrm{e}^{-\lambda(1+\frac{1}{\theta})}\lambda^{k+\alpha-1}\mathrm{d}\lambda.$$

由附录 A 中关于 gamma 分布的定义, 表达式可以写成

$$p_k = \frac{\Gamma(k+\alpha)}{k!\Gamma(\alpha)} \frac{\theta^k}{(1+\theta)^{k+\alpha}} = \binom{k+\alpha-1}{k} \left(\frac{\theta}{1+\theta}\right)^k \left(\frac{1}{1+\theta}\right)^\alpha.$$

这个公式与 (4.9) 式的形式相同, 这表明与 gamma 分布混合的混合 Poisson 分布和负二项分布相同.

值得注意的是, Poisson 分布是负二项分布的一个极限情况. 令 r 趋于无穷大同时 β 趋于零, 并保持它们的乘积不变, 令 $\lambda = r\beta$ 为一常数. 在概率生成函数中做替换 $\beta = \lambda/r$ 得到 (第 3 行和第 5 行用到洛必达法则)

$$\begin{aligned}
\lim_{r\to\infty}\left[1-\frac{\lambda(z-1)}{r}\right]^{-r} &= \exp\left\{\lim_{r\to\infty} -r\ln\left[1-\frac{\lambda(z-1)}{r}\right]\right\} \\
&= \exp\left\{-\lim_{r\to\infty}\frac{\ln[1-\lambda(z-1)/r]}{r^{-1}}\right\} \\
&= \exp\left\{\lim_{r\to\infty}\frac{[1-\lambda(z-1)/r]^{-1}\lambda(z-1)/r^2}{r^{-2}}\right\} \\
&= \exp\left[\lim_{r\to\infty}\frac{r\lambda(z-1)}{r-\lambda(z-1)}\right] \\
&= \exp\{\lim_{r\to\infty}[\lambda(z-1)]\} = \exp[\lambda(z-1)],
\end{aligned}$$

这就是 Poisson 分布的概率生成函数.

4.6.4 二项分布

二项分布也是一类计数分布, 是对索赔数建模时很自然的选择. 它的很多性质与 Poisson 分布和负二项分布不同, 使其具有特殊的用处. 首先, 二项分布的方差小于均值, 因此适用于样本方差小于样本均值的数据集. 与之不同, 负二项分布的方差大于均值, 而 Poisson 分布的方差等于均值.

其次, 它描述了一种由 m 个索赔或损失风险构成的自然状况. 考虑 m 个独立同分布的风险个体, 每个个体的索赔发生概率均为 q. 这适用于寿险的情况, 其中所有的个体都服从相同的死亡表. 例如这个群体可以是 35 岁的男性吸烟者, 已生效 5 年期的保单, 在这种情况下, q 表示个体在下一年死亡的概率. 则每个个体的索赔次数服从 Bernoulli 分布, 取值为 0 的概率为 $1-q$, 取值为 1 的概率为 q. 因此每个人索赔次数的概率生成函数为

$$P(z) = (1-q)z^0 + qz^1 = 1 + q(z-1).$$

如果现在有 m 个独立的个体, 则总索赔次数的概率生成函数是所有个体索赔次数的概率生成函数的乘积, 可表示为

$$P(z) = [1+q(z-1)]^m, \quad 0 < q < 1.$$

由此, 易证恰好发生 k 次索赔的概率是

$$p_k = \Pr(N = k) = \binom{m}{k} q^k (1-q)^{m-k}, \quad k = 0, 1, \cdots, m,$$

得到二项分布的概率函数, 参数为 m 和 q. 在 Bernoulli 情形, 显然群体的最大索赔次数为 m. 因此, 分布仅在不超过 m 的非负整数上有概率.

因此, 二项分布另一个性质是支集有限, 即, 分布取值的范围有限. 这一点非常有用. 比如, 在交通事故中的受伤人数, 或是健康保险中家庭成员个数的建模. 在每种情况下, 可能取值都是有上限的. 在某些情形下, 也可以认为分布在高于某个值数以上的概率很小. 例如, 在对机动车年事故次数建模时, 次数超过某个值如 12 的概率极其小, 因为任何两次事故的间隔还要包括修理的时间. 所以, 如果在某个模型中超过 12 的部分仍然有正的概率, 则概率值必然很小, 所以这些值对于决策的影响甚微. 二项分布的均值和方差如下:

$$\mathrm{E}(N) = mq, \quad \mathrm{Var}(N) = mq(1-q).$$

4.6.5 $(a, b, 0)$ 分布类

下面的定义刻画了此类分布的特点.

定义 4.41 *令 p_k 为某个离散随机变量的概率函数. 称其属于 $(a, b, 0)$ 分布类, 若存在常数 a 和 b 使得下式成立*

$$\frac{p_k}{p_{k-1}} = a + \frac{b}{k}, \quad k = 1, 2, 3, \cdots.$$

这个递推表达式描述了这类计数分布在相邻概率上的大小关系. 实际上, 零点的概率 p_0 可以从递推式的总和为 1 得到, 由此给出了一个边界条件. $(a, b, 0)$ 分布类是一个两参数族, 参数为 a 和 b. 将 Poisson 分布、二项分布、负二项分布的概率函数代入递推公式左边表达式, 可以看出三个分布均满足该递推式, 表 4-1 列出了这些分布的 a 和 b 值同时给出了 p_0 的值, 也列出了几何分布的参数, 这相当于负二项分布的一个参数的特例 $(r = 1)$.

表 4-1 $(a, b, 0)$ 类的分布

分　　布	a	b	p_0
Poisson 分布	0	λ	$\mathrm{e}^{-\lambda}$
二项分布	$-\frac{q}{1-q}$	$(m+1)\frac{q}{1-q}$	$(1-q)^m$
负二项分布	$\frac{\beta}{1+\beta}$	$(r-1)\frac{\beta}{1+\beta}$	$(1+\beta)^{-r}$
几何分布	$\frac{\beta}{1+\beta}$	0	$(1+\beta)^{-1}$

可以证明, 也只有这些分布满足上述的递推式 (见 Panjer and Willmot[106] 第 6 章).

递推公式也可以表示为

$$k\frac{p_k}{p_{k-1}} = ak + b, \quad k = 1, 2, 3, \cdots.$$

表达式的左边为 k 的线性函数. 由表 4-1, Poisson 分布的斜率 a 为 0, 二项分布的斜率为负值, 而负二项分布包括几何分布的斜率为正值. 这里给出了一种模型拟合时选择分布的比较直观的图形方法. 首先, 可以按照下面的近似公式绘出关于 k 的图形

$$k\frac{\hat{p}_k}{\hat{p}_{k-1}} = k\frac{n_k}{n_{k-1}}.$$

若选择的模型适当, 由观测值近似得出的应该是一条直线, 直线的斜率应该表示适用的模型, 但是应该注意到只要某个 $n_k = 0$ 则该方法不可行. 因此这种方法对于观测量不大的数据不太适用.

例 4.42 *参照表 4-2 给出的事故数据 (来源于 Thyrion[128]), 总共有 9 461 个机动车事故保单. 线性部分的数值也在表中给出.*

表 4-2 事故数据

事故发生数, k	保单数, n_k	$k\frac{n_k}{n_{k-1}}$
0	7 840	
1	1 317	0.17
2	239	0.36
3	42	0.53
4	14	1.33
5	4	1.43
6	4	6.00
7	1	1.75
8+	0	
总和	9 461	

图 4-8 为事故次数 k 与比值之间的图像, 可以看出除了 $k = 6$ 这个点比值近似呈一条直线. 因为观测值的个数随 k 的增加而减少, 因此比值的可靠性变小同时可变性增加. 这说明了这种特殊方法的不足之处. 直观上, 所有的点都应该有相同的价值, 然而, 左边的点比右边的点更加可靠. 由图中可以看出, 数据近似呈直线, 而直线的斜率是正的. 这说明负二项分布可以作为一个合适的模型. 而斜率是否远离 0 并不能从图中明显看出. 如果比例放大纵轴, 直线可能会看起来更陡, 因此斜率更明显地不为 0. 但是, 从图上很难区分是 Poisson 分布还是负二项分布, 因为

Poisson 分布斜率为零. 然而因为斜率非负可以推断二项分布并不是好的选择. 在这种情况下, 建议选用 Poisson 分布和负二项分布, 进而再进行正规的统计检验从中做出选择.

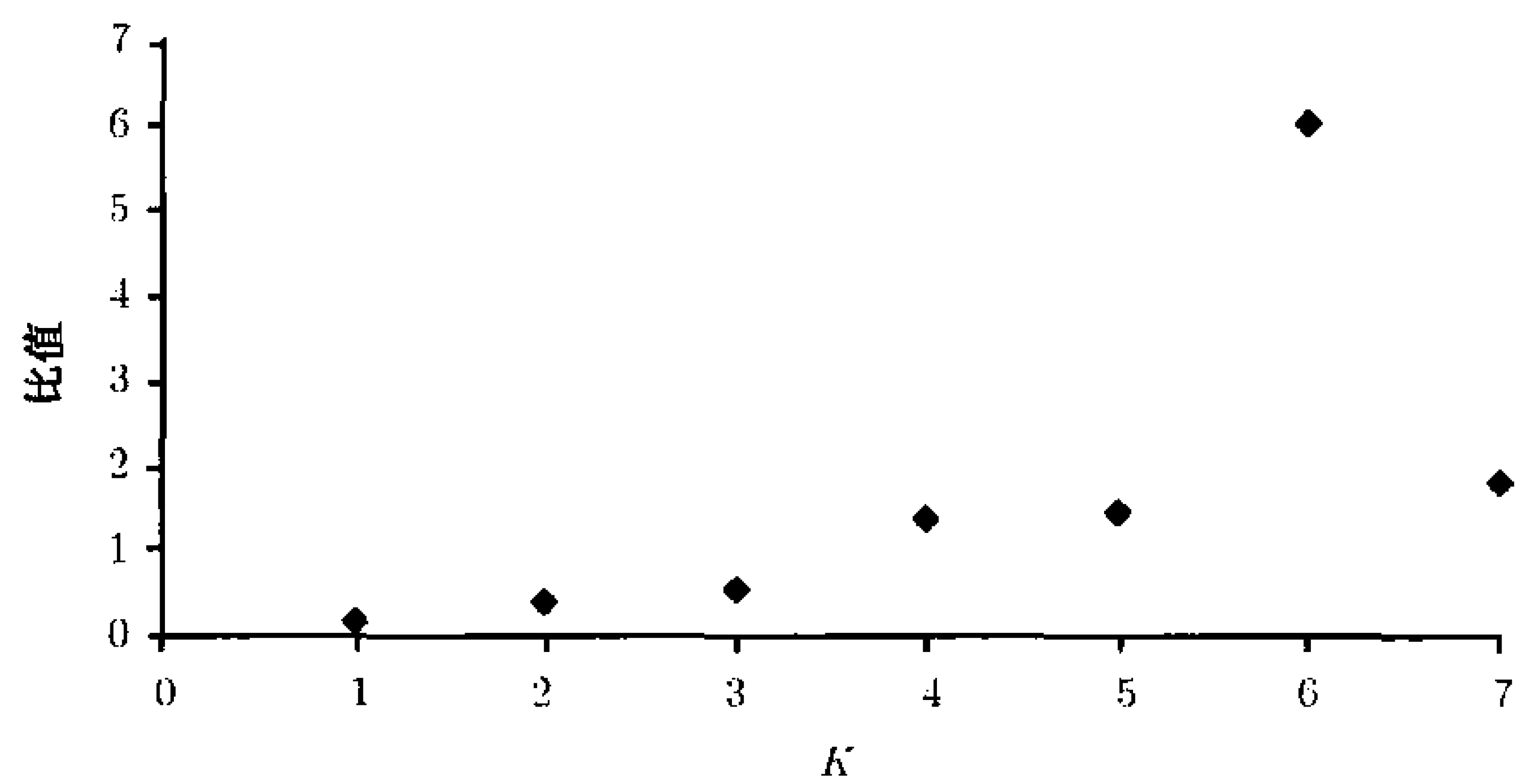


图 4-8

另一种方法是通过观测均值和方差的关系比较分布的近似程度. 在这个数据集中, 索赔次数的均值为 0.214 4, 方差为 0.288 9. 因为方差高于均值, 负二项分布要比 Poisson 分布更合适. 这又存在一个量的问题, 没有一个标准的方法来检验是否方差足够大于均值, 进而选择负二项分布. 为了考虑更为正规的分析, 表 4-3 给出了 Poisson 分布和负二项分布的最大似然参数估计 (第 12 章介绍) 和每种情况下的对数似然值的负值. 第 13 章将提出正规的模型选择方法. 这些方法表明对于该数据集, 负二项分布优于 Poisson 分布. 然而这些方法也表明负二项分布也不是一个最好的模型, 因此可以考虑引进一些其他分布.

表 4-3 Poisson 分布和负二项分布的比较

分　布	参数估计	一对数似然函数值
Poisson 分布	$\hat{\lambda} = 0.214\ 353\ 7$	5 490.78
负二项分布	$\hat{\beta} = 0.305\ 559\ 4$ $\hat{r} = 0.701\ 512\ 2$	5 348.04

在 4.6.6 节, 我们会讨论比 Poisson 分布、二项分布和负二项分布更一般的模型, 从而扩充这三个分布.

4.6.6 分布在零点的截断和修正

有时候, 前面讨论的那些分布无法充分地描述实践中遇到的数据集的特性. 这也许是因为负二项分布尾部的厚度不够, 或是因为 $(a, b, 0)$ 分布类不能刻画数据集在某个特殊部分的性状.

本节要考虑在左端点特别是零点的拟合效果较差的问题.

对于保险计数数据, 在零点的概率表示在观察期内没有索赔发生的概率, 因为

在保险中损失的发生概率一般都是很低的, 因此在零点有较大的概率值. 故对于这个点的拟合值应该给予特别的关注.

还有一些情况自然会在零点出现相当大的概率. 考虑团体牙医保险, 如果某个家庭中的丈夫和妻子各自都参加了其雇主的计划, 两个团体保险的契约都将提供家庭成员的保障, 那么, 夫妻将选择两个计划中受益更好的保险人的赔付, 因此, 另一个契约则不会有索赔发生. 所以, 在某个保险人的研究中, 可以发现较高的预期无索赔数.

类似地, 也可能发现在零点的概率低于期望数值甚至是零的情形. 例如, 如果记录事故导致的索赔数, 其最小的观测值是 1.

对于 Poisson 分布、二项分布和负二项分布, 在零点处的调整很容易操作.

定义 4.43 *令 p_k 为某个离散随机变量的概率函数. 称其属于 $(a,b,1)$ 分布类, 若存在常数 a 和 b, 使得下式成立*

$$\frac{p_k}{p_{k-1}}=a+\frac{b}{k},\quad k=2,3,4,\cdots.$$

注意到它与 $(a,b,0)$ 分布类的唯一区别是递推从 p_1 而非 p_0 开始的. 这使得该分布从 $k=1$ 到 $k=\infty$ 处与 $(a,b,0)$ 分布类在概率的和是一个常数的情况下的形状相同, 因为 $\sum\limits_{k=1}^{\infty}p_k$ 可以设定为 (0, 1] 中的任何一个值, 剩下的即为 $k=0$ 的概率.

将 $p_0=0$ 和 $p_0>0$ 的情况分开讨论. 第一类称为**截断**(更特别地, 称为**零点截断**) 分布. 其成员有零点截断 Poisson 分布、零点截断二项分布和零点截断负二项分布 (包括它的特例, 零点截断几何分布).

第二类称为**零点修正**(简称 ZM) 分布, 因为其概率是由 $(a,b,0)$ 分布类修正而得到的. 这类分布可以被看成是由 $(a,b,0)$ 分布类和一个在零点处的退化分布混合而成. 或者也可以称之为**含零点的截断分布**, 因为可看成是一个截断分布和一个零点退化分布的混合. 下面将进行正式的描述. 同时零点截断分布也可以看成是 $p_0=0$ 的零点修正分布.

在三种分布中, 符号可能会混淆. 当表示一般的离散分布时, 我们将继续令 $p_k=\Pr(N=k)$, 用 p_k^T 表示零点截断分布, 用 p_k^M 表示零点修正分布. 再明确一下, 零点修正分布也可以是一个零点截断分布.

令 $P(z)=\sum\limits_{k=0}^{\infty}p_kz^k$ 表示 $(a,b,0)$ 分布类的概率生成函数, 令 $P^M(z)=\sum\limits_{k=0}^{\infty}p_k^Mz^k$ 表示其对应的 $(a,b,1)$ 分布类的概率生成函数. 即

$$p_k^M=cp_k,\quad k=1,2,3,\cdots,$$

其中 p_0^M 为任意 (0, 1) 内的实数. 则

$$\begin{aligned}P^M(z) &= p_0^M + \sum_{k=1}^{\infty} p_k^M z^k = p_0^M + c\sum_{k=1}^{\infty} p_k z^k \\ &= p_0^M + c[P(z) - p_0].\end{aligned}$$

因为 $P^M(1) = P(1) = 1$, 并且

$$1 = p_0^M + c(1 - p_0),$$

结果为

$$c = \frac{1 - p_0^M}{1 - p_0} \text{ 或 } p_0^M = 1 - c(1 - p_0).$$

应保证 p_k^M 的和为 1, 我们有

$$P^M(z) = p_0^M + \frac{1 - p_0^M}{1 - p_0}[p(z) - p_0] = \left(1 - \frac{1 - p_0^M}{1 - p_0}\right)1 + \frac{1 - p_0^M}{1 - p_0}P(z). \tag{4.10}$$

这是某个退化分布和 $(a, b, 0)$ 分布类的概率生成函数的加权平均. 进一步有

$$p_k^M = \frac{1 - p_0^M}{1 - p_0}p_k, \quad k = 1, 2, \cdots. \tag{4.11}$$

令 $P^T(z)$ 表示某个 $(a, b, 0)$ 分布类的概率生成函数 $P(z)$ 所对应的零点截断分布的概率生成函数. 则在 (4.10) 和 (4.11) 中取 $p_0^M = 0$, 有

$$P^T(z) = \frac{P(z) - p_0}{1 - p_0},$$

$$p_k^T = \frac{p_k}{1 - p_0}, \quad k = 1, 2, \cdots. \tag{4.12}$$

则由 (4.11) 式, 有

$$p_k^M = (1 - p_0^M)p_k^T, \quad k = 1, 2, \cdots, \tag{4.13}$$

$$P^M(z) = p_0^M(1) + (1 - p_0^M)p^T(z). \tag{4.14}$$

因此, 零点修正分布也是某个退化分布和 $(a, b, 0)$ 分布类零点截断分布的加权平均. 下面的例子将具体说明这些关系.

例 4.44 考虑某个参数为 $\beta = 0.5, r = 2.5$ 的负二项分布随机变量. 确定该随机变量在前 4 个点上的概率. 并给出零点截断和零点修正 (已知 $p_0^M = 0.6$) 形式下的概率.

解 由 4.6.7 节的表 4-4, 对于负二项分布, 有

$$p_0=(1+0.5)^{-2.5}=0.362\ 887,$$
$$a=\frac{0.5}{1.5}=\frac{1}{3},$$
$$b=\frac{(2.5-1)(0.5)}{1.5}=\frac{1}{2}.$$

前 3 个递推式为

$$p_1=0.362\ 887\left(\frac{1}{3}+\frac{1}{2}\frac{1}{1}\right)=0.302\ 406,$$
$$p_2=0.302\ 406\left(\frac{1}{3}+\frac{1}{2}\frac{1}{2}\right)=0.176\ 404,$$
$$p_3=0.176\ 404\left(\frac{1}{3}+\frac{1}{2}\frac{1}{3}\right)=0.088\ 202.$$

对于零点截断的随机变量, 由定义 $p_0^T=0$. 由第一个递推式 (4.12) 式, 有 $p_1^T=0.302\ 406/(1-0.362\ 887)=0.474\ 651$. 则

$$p_2^T=0.474\ 651\left(\frac{1}{3}+\frac{1}{2}\frac{1}{2}\right)=0.276\ 880,$$
$$p_3^T=0.276\ 880\left(\frac{1}{3}+\frac{1}{2}\frac{1}{3}\right)=0.138\ 440.$$

如果最初的值均已知, 则零点截断后的概率可由原始值乘上 $1/(1-0.362\,887)=1.569\ 580$ 得到.

对于零点修正随机变量, 可以令 $p_0^M=0.6$. 由 (4.11) 式, 有 $p_1^M=(1-0.6)(0.302\ 406)/(1-0.362\ 887)=0.189\ 860$. 则

$$p_2^M=0.189\ 860\left(\frac{1}{3}+\frac{1}{2}\frac{1}{2}\right)=0.110\ 752,$$
$$p_3^M=0.110\ 752\left(\frac{1}{3}+\frac{1}{2}\frac{1}{3}\right)=0.055\ 376.$$

这样, 原始的负二项分布概率都将乘上 $(1-0.6)/(1-0.362\ 887)=0.627\ 832$. 注意到, 对于所有的 $j\geqslant 1$, 有 $p_j^M=0.4p_j^T$. □

尽管, 只讨论了 $(a,b,0)$ 分布类的零点修正分布, 但 $(a,b,1)$ 分布类远不止如此, (a,b) 参数空间可以扩展到包括负二项分布 $-1<r<0$ 的情况. 对于 $(a,b,0)$ 分布类, 必须要求 $r>0$. 增加新的区域后, 扩展的截断负二项分布 (ETNB) 的参数约束为 $\beta>0,r>-1,r\neq 0$.

为了证明递推式

$$p_k=p_{k-1}\left(a+\frac{b}{k}\right),\quad k=2,3,\cdots, \tag{4.15}$$

当 $p_0=0$ 时定义了一个先验分布, 可以证明对于任何 p_1 的取值, 由递推式得到的 p_k 都是正的, 并且有 $\sum_{k=1}^{\infty} p_k<\infty$. 对于 ETNB, 参数空间必须满足

$$\begin{aligned} a &= \frac{\beta}{1+\beta}, \quad \beta>0, \\ b &= (r-1)\frac{\beta}{1+\beta}k, \quad r>-1, r\neq 0 \end{aligned}$$

(见习题 4.44).

当 $r\to 0$ 时, ETNB 的极限情况是对数分布, 其中

$$p_k^T=\frac{[\beta/(1+\beta)]^k}{k\ln(1+\beta)}, \quad k=1,2,3,\cdots \tag{4.16}$$

(见习题 4.45). 对数分布的概率生成函数为

$$P^T(z)=1-\frac{\ln[1-\beta(z-1)]}{\ln(1+\beta)} \tag{4.17}$$

(见习题 4.46). 零点修正的对数分布可以通过以下方法生成. 任意给定零点的概率, 然后相应的整体减少其他点的概率值.

一个有趣的情况是, $-1<r<0, \beta\to\infty$ 时的特殊的极限情况是一个分布, 有时称为 Sibuya 分布. 其概率生成函数为 $P(z)=1-(1-z)^{-r}$ 并且矩函数不存在 (见习题 4.47). 不存在矩函数的分布通常不用作索赔次数的模型 (除非右尾部被修正), 因为这种分布的期望索赔次数为无穷大, 这样很难进行定价.

例 4.45　确定参数为 $r=-0.5, \beta=1$ 的 ETNB 分布的概率. 在截断情形和 $p_0^M=0.6$ 的修正情形分别给出其结果.

解　我们有 $a=1/(1+1)=0.5$ 和 $b=(-0.5-1)(1)/(1+1)=-0.75$. 由附录 B 得到 $p_1^T=-0.5(1)/[(1+1)^{0.5}-(1+1)]=0.853\ 553$. 接下来的值为

$$\begin{aligned} p_2^T &= \left(0.5-\frac{0.75}{2}\right)(0.853\ 553)=0.106\ 694, \\ p_3^T &= \left(0.5-\frac{0.75}{3}\right)(0.106\ 694)=0.026\ 674. \end{aligned}$$

对于修正概率, 截断概率需要乘上 0.4 得到 $p_1^M=0.341\ 421$, $p_2^M=0.042\ 678$, $p_3^M=0.010\ 670$.

一个合理的问题是 ETNB 分布类是否存在一个 “天然” 的分布, 即它的递推式从 p_1 开始而不从 p_2 开始, 为此, p_0 必须满足 $p_1=(0.5-0.75/1)p_0=-0.25p_0$. 这需要两个概率中至少一个是负值, 因此无解. 易于证明, 当 $r<0$ 时, 将出现这种结果. □

除了上面讨论的分布, $(a,b,1)$ 分布类不再包含其他的分布, 表 4-4 为总结.

表 4-4 $(a, b, 1)$ 分布类

分布[a]	p_0	a	b	参数空间
Poisson	$e^{-\lambda}$	0	λ	$\lambda > 0$
ZT Poisson	0	0	λ	$\lambda > 0$
ZM Poisson	任意的	0	λ	$\lambda > 0$
二项	$(1-q)^m$	$-\frac{q}{1-q}$	$(m+1)\frac{q}{1-q}$	$0 < q < 1$
ZT 二项	0	$-\frac{q}{1-q}$	$(m+1)\frac{q}{1-q}$	$0 < q < 1$
ZM 二项	任意的	$-\frac{q}{1-q}$	$(m+1)\frac{q}{1-q}$	$0 < q < 1$
负二项	$(1+\beta)^{-r}$	$\frac{\beta}{1+\beta}$	$(r-1)\frac{\beta}{1+\beta}$	$r > 0, \beta > 0$
ETNB	0	$\frac{\beta}{1+\beta}$	$(r-1)\frac{\beta}{1+\beta}$	$r > -1$,[b] $\beta > 0$
ZM ETNB	任意的	$\frac{\beta}{1+\beta}$	$(r-1)\frac{\beta}{1+\beta}$	$r > -1$,[b] $\beta > 0$
几何	$(1+\beta)^{-1}$	$\frac{\beta}{1+\beta}$	0	$\beta > 0$
ZT 几何	0	$\frac{\beta}{1+\beta}$	0	$\beta > 0$
ZM 几何	任意的	$\frac{\beta}{1+\beta}$	0	$\beta > 0$
对数	0	$\frac{\beta}{1+\beta}$	$-\frac{\beta}{1+\beta}$	$\beta > 0$
ZM 对数	任意的	$\frac{\beta}{1+\beta}$	$-\frac{\beta}{1+\beta}$	$\beta > 0$

a ZT= 零点截断, ZM= 零点修正.

b 除 $r = 0$ 外, 均为对数分布.

4.6.7 频率的复合模型

更大的一类频率分布族可以是将两种离散分布进行复合. **复合**一词表示新分布的概率生成函数 $P(z)$ 可以写成

$$P(z) = P_N[P_M(z)], \tag{4.18}$$

其中 $P_N(z)$ 和 $P_M(z)$ 分别称为**主分布**和**次分布**.

复合分布是通过如下的方法自然产生的. 令 N 表示计数随机变量, 概率生成函数为 $P_N(z)$. 令 $M_1, M_2, \cdots$ 是独立同分布的随机变量, 概率生成函数均为 $P_M(z)$. 假设所有的 M_j 都不依赖于 N, 随机变量和 $S = M_1 + M_2 + \cdots + M_N$ 的概率生成函数 ($N = 0$ 时 $S = 0$) 为 $P_S(z) = P_N[P_M(z)]$. 表达式如下

$$P_S(z) = \sum_{k=0}^{\infty} \Pr(S = k) z^k = \sum_{k=0}^{\infty} \sum_{n=0}^{\infty} \Pr(S = k | N = n) \Pr(N = n) z^k$$

$$=\sum_{n=0}^{\infty}\Pr(N=n)\sum_{k=0}^{\infty}\Pr(M_1+\cdots+M_n=k|N=n)z^k$$
$$=\sum_{n=0}^{\infty}\Pr(N=n)[P_M(z)]^n=P_N[P_M(z)].$$

保险契约会自然形成这种分布. 如果 N 代表风险投资组合中事故发生的次数, $\{M_k;k=1,2,\cdots,N\}$ 表示某个承保事故中的索赔次数 (受伤的个数或车辆数), S 表示保单组合的总索赔次数. 当然, 这种解释对于证明复合分布的作用并不是必要的, 如果复合分布能够很好地拟和数据, 就足以证明其本身是合适的. 在 4.6.12 节中还介绍了一些其他运用这种分布的情形.

例 4.46 证明任何零点修正分布是一个复合分布.

证明 考虑主分布为 Bernoulli 分布, 概率生成函数为 $P_N(z)=1-q+qz$. 再考虑任意一个概率生成函数为 $P_M(z)$ 的次分布. 则由 (4.18) 式得到

$$P_S(z)=P_N[P_M(z)]=1-q+qP_M(z).$$

由 (4.10) 式得知这是一个 ZM 分布的概率生成函数, 其中

$$q=\frac{1-p_0^M}{1-p_0}.$$

即 ZM 分布在零点分配了任意的概率值 p_0^M, 而 p_0 是次分布在零点的概率. □

例 4.47 考虑 M 和 N 都是 Poisson 分布的情况. 确定其分布的概率生成函数.

解 称这类分布为 Poisson-Poisson 分布或 A 类 Neyman 分布. 令 $P_N(z)=\mathrm{e}^{\lambda_1(z-1)}$, $P_M(z)=\mathrm{e}^{\lambda_2(z-1)}$, 则

$$P_S(z)=\mathrm{e}^{\lambda_1[\mathrm{e}^{\lambda_2(z-1)}-1]}.$$

当 λ_2 比 λ_1 大很多时, 比如 $\lambda_1=0.1$ 和 $\lambda_2=10$, 则产生的分布具有两个局部峰值点. □

保单组合恰好发生 k 次索赔的概率可以表示为

$$\begin{aligned}\Pr(S=k)&=\sum_{n=0}^{\infty}\Pr(S=k|N=n)\Pr(N=n)\\&=\sum_{n=0}^{\infty}\Pr(M_1+\cdots+M_N=k|N=n)\Pr(N=n)\\&=\sum_{n=0}^{\infty}\Pr(M_1+\cdots+M_n=k)\Pr(N=n).\end{aligned}\tag{4.19}$$

令 $g_n=\Pr(S=n)$, $p_n=\Pr(N=n)$ 和 $f_n=\Pr(M=n)$, 还可以表示为

$$g_k=\sum_{n=0}^{\infty}p_nf_k^{*n},\tag{4.20}$$

其中 $f_k^{*n}, k=0,1,\cdots$, 是函数 f_k 的 “n 重卷积”, 即 n 个概率函数为 f_k 的独立同分布 (i.i.d.) 随机变量之和等于 k 的概率.

若 $P_N(z)$ 是取自 $(a,b,0)$ 分布类,

$$p_k=\left(a+\frac{b}{k}\right)p_{k-1},\quad k=1,2,\cdots, \tag{4.21}$$

则可以用一个简单的递推公式进行计算. 这个公式避免了卷积计算从而减少了相当多的计算量.

定理 4.48 如果主分布为 $(a,b,0)$ 分布类, 则有如下的递推公式成立:

$$g_k=\frac{1}{1-af_0}\sum_{j=1}^{k}\left(a+\frac{b_j}{k}\right)f_jg_{k-j},\quad k=1,2,3,\cdots. \tag{4.22}$$

证明 由 (4.21) 式,

$$np_n=a(n-1)p_{n-1}+(a+b)p_{n-1}.$$

两边同时乘上 $[P_M(z)]^{n-1}P_M'(z)$, 并对 n 个式子求和,

$$\sum_{n=1}^{\infty}np_n[P_M(z)]^{n-1}P_M'(z)=a\sum_{n=1}^{\infty}(n-1)p_{n-1}[P_M(z)]^{n-1}P_M'(z)$$
$$+(a+b)\sum_{n=1}^{\infty}p_{n-1}[P_M(z)]^{n-1}P_M'(z).$$

因为 $P_S(z)=\sum\limits_{n=0}^{\infty}p_n[P_M(z)]^n$, 利用前面的等式有

$$P_S'(z)=a\sum_{n=0}^{\infty}np_n[P_M(z)]^nP_M'(z)+(a+b)\sum_{n=0}^{\infty}p_n[P_M(z)]^nP_M'(z).$$

因此

$$P_S'(z)=aP_S'(z)P_M(z)+(a+b)P_S(z)P_M'(z).$$

两边都可以展开为 z 的幂形式, 等式两边 z^{k-1} 的系数必然要相等. 因此, 对于 $k=1,2,\cdots$, 有

$$\begin{aligned}kg_k&=a\sum_{j=0}^{k}(k-j)f_jg_{k-j}+(a+b)\sum_{j=0}^{k}jf_jg_{k-j}\\&=akf_0g_k+a\sum_{j=1}^{k}(k-j)f_jg_{k-j}+(a+b)\sum_{j=1}^{k}jf_jg_{k-j}\\&=akf_0g_k+ak\sum_{j=1}^{k}f_jg_{k-j}+b\sum_{j=1}^{k}jf_jg_{k-j}.\end{aligned}$$

因此

$$g_k = af_0g_k + \sum\nolimits_{j=1}^{k}\left(a+\frac{b_j}{k}\right)f_jg_{k-j}.$$

适当整理上式即为 (4.22) 式. □

为了利用 (4.22) 式, 需要给出初始值 g_0, 如定理 4.51. 如果主分布是 $(a,b,1)$ 分布类, 则需要适当的修正以前的证明以使主分布的递推从 $k=2$ 开始. 结果如下.

定理 4.49 如果主分布是 $(a,b,1)$ 分布类, 则递推公式为

$$g_k = \frac{[p_1-(a+b)p_0]f_k + \sum_{j=1}^{k}(a+bj/k)f_jg_{k-j}}{1-af_0}, \quad k=1,2,3,\cdots. \tag{4.23}$$

证明类似于定理 4.48, 留给读者完成.

例 4.50 若主分布为 Poisson 分布, 推导递推公式.

解 在这种情况下有 $a=0, b=\lambda$, 得到的递推式为

$$g_k = \frac{\lambda}{k}\sum_{j=1}^{k} jf_jg_{k-j}.$$

由 (4.18) 式, 初始值为

$$g_0 = \Pr(S=0) = P(0) = P_N[P_M(0)] = P_N(f_0) = \mathrm{e}^{-\lambda(1-f_0)}. \tag{4.24}$$

称这类分布为复合 Poisson 分布. 当次分布确定后, 复合分布称为 Poisson-X 分布, X 是次分布的名称. □

上述获得 g_0 的方法也适用于任何复合分布.

定理 4.51 对于任何复合分布, $g_0 = P_N(f_0)$, 其中 $P_N(z)$ 是主分布的概率生成函数, f_0 是次分布在零点的概率.

证明 见 (4.24) 式第 3 个和第 4 个等式. □

应该注意的是次分布并不需要什么特殊的形式, 然而, 为了适当限制分布的范围, 次分布通常选自 $(a,b,0)$ 分布类或是 $(a,b,1)$ 分布类.

例 4.52 计算 Poisson-ETNB 分布的概率, 其中 Poisson 分布的参数为 $\lambda=3$, ETNB 分布的参数为 $r=-0.5$ 和 $\beta=1$.

解 由例 4.45 知, 次分布的概率为 $f_0=0$, $f_1=0.853\ 553$, $f_2=0.106\ 694$ 和 $f_3=0.026\ 674$. 由 (4.24) 式, $g_0=\exp[-3(1-0)]=0.049\ 787$. 对于 Poisson 主分布, $a=0, b=3$. 则 (4.22) 递推式为

$$g_k = \frac{\sum_{j=1}^{k}(3j/k)f_jg_{k-j}}{1-0(0)} = \sum_{j=1}^{k}\frac{3j}{k}f_jg_{k-j}.$$

则

$$g_1=\frac{3(1)}{1}0.853\,553(0.049\,787)=0.127\,488,$$

$$g_2=\frac{3(1)}{2}0.853\,553(0.127\,488)+\frac{3(2)}{2}0.106\,694(0.049\,787)=0.179\,163,$$

$$g_3=\frac{3(1)}{3}0.853\,553(0.179\,163)+\frac{3(2)}{3}0.106\,694(0.127\,448)$$
$$+\frac{3(3)}{3}0.026\,674(0.049\,787)=0.184\,114.$$

□

例 4.53 证明 Poisson–对数分布为负二项分布.

证明 负二项分布的概率生成函数为

$$P(z)=[1-\beta(z-1)]^{-r}.$$

假设 $P_N(z)$ 为 Poisson(λ), $P_M(z)$ 为 logarithmic(β) 分布; 则有

$$P_N[P_M(z)]=\exp\{\lambda[P_M(z)-1]\}=\exp\left\{\lambda\left[1-\frac{\ln[1-\beta(z-1)]}{\ln(1+\beta)}-1\right]\right\}$$
$$=\exp\left\{\frac{-\lambda}{\ln(1+\beta)}\ln[1-\beta(z-1)]\right\}$$
$$=[1-\beta(z-1)]^{-\lambda/\ln(1+\beta)}=[1-\beta(z-1)]^{-r},$$

其中 $r=\lambda/\ln(1+\beta)$. 这说明负二项分布可以写成 Poisson 分布与对数次分布的复合分布.

□

上面的例子说明 Poisson–对数分布的构造并没有超出 $(a,b,0)$ 分布类或是 $(a,b,1)$ 分布类. 因此, 这样的分布组合对我们没有用处. 另一个分布组合未超出 $(a,b,1)$ 分布类的例子是: 主分布和次分布均为几何分布, 其结果为零点修正几何分布, 见习题 4.51. 定理 4.5.4 将证明某些分布的复合无法构造出超出原有分布类的分布. 如前所设 $P_S(z)=P_N[P_M(z);\theta]$, 现在, $P_M(z)$ 总是可以表示为

$$P_M(z)=f_0+(1-f_0)P_M^*(z), \tag{4.25}$$

其中 $P_M^*(z)$ 是在正值范围 (也可以说是零点截断类) 条件下分布的概率生成函数.

定理 4.54 设概率生成函数为 $P_N(z;\theta)$ 满足

$$P_N(z;\theta)=B[\theta(z-1)].$$

对某个参数 θ 和某个与 θ 独立的函数 $B(z)$, 即参数 θ 和变量 z 仅以 $\theta(z-1)$ 的形式出现在概率生成函数中, 当然也许还有其他参数出现在概率生成函数中. 则 $P_S(z)=P_N[P_M(z);\theta]$ 可以表示为

$$P_S(z)=P_N[P_M^T(z);\theta(1-f_0)].$$

证明

$$\begin{aligned}P_S(z)&=P_N[P_M(z);\theta]=P_N[f_0+(1-f_0)P_M^T(z);\theta]\\&=B\{\theta[f_0+(1-f_0)P_M^T(z)-1]\}\\&=B\{\theta(1-f_0)[P_M^T(z)-1]\}\\&=P_N[P_M^T(z);\theta(1-f_0)].\end{aligned}$$

□

这说明增加、删除或修正次分布在零点的概率并不能构造出新的分布, 因为这等价于对主分布的参数 θ 进行修正. 这表明, 例如, Poisson 主分布与某个 Poisson 分布复合、零点截断 Poisson 分布或是零点修正 Poisson 分布都只能产生 A 类 Neyman(Poisson-Poisson) 分布.

例 4.55　确定 Poisson–零点修正 ETNB 分布的概率, 其中参数为 $\lambda=7.5$, $p_0^M=0.6$, $r=-0.5$ 和 $\beta=1$.

解　由例 4.45 知, 次分布的概率为 $f_0=0.6$, $f_1=0.341\ 421$, $f_2=0.042\ 678$ 和 $f_3=0.010\ 670$. 由 (4.24) 式, $g_0=\exp[-7.5(1-0.6)]=0.049\ 787$. 对于 Poisson 主分布, $a=0, b=7.5$. 则递推式 (4.22) 为

$$g_k=\frac{\sum_{j=1}^{k}(7.5j/k)f_jg_{k-j}}{1-0(0.6)}=\sum_{j=1}^{k}\frac{7.5j}{k}f_jg_{k-j}.$$

则

$$\begin{aligned}g_1&=\frac{7.5(1)}{1}0.341\ 421(0.049\ 787)=0.127\ 487,\\g_2&=\frac{7.5(1)}{2}0.341\ 421(0.127\ 487)+\frac{7.5(2)}{2}0.042\ 678(0.049\ 787)=0.179\ 161,\\g_3&=\frac{7.5(1)}{3}0.341\ 421(0.179\ 161)+\frac{7.5(2)}{3}0.042\ 678(0.127\ 487)\\&\quad+\frac{7.5(3)}{3}0.010\ 670(0.049\ 787)=0.184\ 112.\end{aligned}$$

除了一些细微的差别, 这些概率与例 4.52 得到的概率相同. □

4.6.8　复合 Poisson 分布族的性质

频率复合模型中最重要的一类是复合 Poisson 频率分布. 采用这个模型的一个自然的考虑是 Poisson 分布可以很好地描述造成索赔的事故次数, 而一次事故导致的赔案数本身也是随机变量, 这一点在前面的小节中已有讨论. 然而复合 Poisson 分布族还有很多很好的数学性质, 特别地, 如前面小节中讨论的概率计算递推公式. 另外, 在 4.6.10 节中还会更详细地讨论复合 Poisson 分布与混合 Poisson 频率分布

的关系. 这里考虑这些分布的其他一些性质. 复合 Poisson 分布的概率生成函数可以表示为

$$P(z)=\exp\{\lambda[Q(z)-1]\}, \tag{4.26}$$

其中 $Q(z)$ 是次分布的概率生成函数.

例 4.56 确定 Poisson-ETNB 分布的概率生成函数, 并证明其类似于 Poisson-负二项分布的概率生成函数.

证明 ETNB 分布的概率生成函数为

$$Q(z)=\frac{[1-\beta(z-1)]^{-r}-(1+\beta)^{-r}}{1-(1+\beta)^{-r}},$$

其中 $\beta>0, r>-1, r\neq 0$. 则 Poisson-ETNB 分布的概率生成函数的对数为

$$\begin{aligned}\ln P(z)&=\lambda\left\{\frac{[1-\beta(z-1)]^{-r}-(1+\beta)^{-r}}{1-(1+\beta)^{-r}}-1\right\}\\&=\lambda\{\frac{[1-\beta(z-1)]^{-r}-1}{1-(1+\beta)^{-r}}\}\\&=\mu\{[1-\beta(z-1)]^{-r}-1\},\end{aligned}$$

其中 $\mu=\lambda/[1-(1+\beta)^{-r}]$. 这样就定义了一个复合 Poisson 分布, 主分布的均值为 μ, 次分布的概率生成函数为 $[1-\beta(z-1)]^{-r}$, 这是一个负二项随机变量的概率生成函数, 只要 r 或 μ 为正值. 这说明次分布在零点的概率对于复合 Poisson 形式没有影响. 上面的计算 $\ln P(z)=\mu\{[1-\beta(z-1)]^{-r}-1\}$ 证明了 Poisson-ETNB 分布的概率生成函数 $P(z)$ 的参数域为 $\{\beta>0, r>-1, \mu r>0\}$, 这一点对于参数估计与分析非常有价值. □

我们可以比较这些分布的峰度 (三阶矩), 产生由峰度值表示的量, 进而刻画分布的尾部, 即使这些分布的均值和方差都是固定的, 这个量也可能不同. 由 (4.26) 式 (见习题 4.53) 和定义 3.4, 复合 Poisson 分布的均值以及二阶中心矩和三阶中心矩为

$$\mu_1'=\mu=\lambda m_1', \quad \mu_2=\sigma^2=\lambda m_2', \quad \mu_3=\lambda m_3', \tag{4.27}$$

其中 m_j' 为次分布的 j 阶原点矩. 峰度系数为

$$\gamma_1=\frac{\mu_3}{\sigma^3}=\frac{m_3'}{\lambda^{1/2}(m_2')^{3/2}}.$$

对于 Poisson–二项分布, 通过简单的代数推导 (见习题 4.54), 得到

$$\mu=\lambda mq, \quad \sigma^2=\mu[1+(m-1)q], \quad \mu_3=3\sigma^2-2\mu+\frac{m-2}{m-1}\frac{(\sigma^2-\mu)^2}{\mu}. \tag{4.28}$$

对于负二项分布、Polya-Aeppli 分布、A 类 Neyman 分布和 Poisson–ETNB 分布也进行类似的练习, 可以得到

$$\text{负二项分布:}\quad \mu_3 = 3\sigma^2 - 2\mu + 2\frac{(\sigma^2-\mu)^2}{\mu},$$

$$\text{Polya-Aeppli 分布:}\quad \mu_3 = 3\sigma^2 - 2\mu + \frac{3}{2}\frac{(\sigma^2-\mu)^2}{\mu},$$

$$\text{A 类 Neyman 分布:}\quad \mu_3 = 3\sigma^2 - 2\mu + \frac{(\sigma^2-\mu)^2}{\mu},$$

$$\text{Poisson-ETNB 分布:}\quad \mu_3 = 3\sigma^2 - 2\mu + \frac{r+2}{r+1}\frac{(\sigma^2-\mu)^2}{\mu}.$$

对于 Poisson–ETNB 分布, r 的取值范围是 $-1 < r < \infty, r \neq 0$. 注意到 $r \to 0$ 时次分布为对数形式, 结果是负二项分布.

注意到对于固定的均值和方差, 5 个分布的三阶矩的区别只来源于最后一项的系数. Poisson 分布满足 $\mu_3 = \lambda = 3\sigma^2 - 2\mu$, 因此每个表达式中 μ_3 的第三项反映了它们与 Poisson 分布的区别. 对于 Poisson–二项分布, 如果 $m = 1$, 则还是 Poisson 分布, 因为其等价于零点截断的 Poisson–二项分布, 而该二项分布退化为在零点的概率为 1. 另一种方法是由 (4.28) 式, 有

$$\begin{aligned}\mu_3 &= 3\sigma^2 - 2\mu + \frac{m-2}{m-1}\frac{(m-1)^2q^4\lambda^2m^2}{\lambda mq}\\ &= 3\sigma^2 - 2\mu + (m-2)(m-1)q^3\lambda m,\end{aligned}$$

当 $m = 1$ 时退化为 Poisson 分布. 因此构造非 Poisson 分布的必要条件是 $m \geqslant 2$, 则系数满足

$$0 \leqslant \frac{m-2}{m-1} < 1.$$

对于 Poisson-ETNB, 因为 $r > -1$, 则系数满足

$$1 < \frac{r+2}{r+1} < \infty,$$

注意当 $r = 0$ 时为负二项分布. 对于 A 类 Neyman 分布, 其系数为 1. 因此这三个分布可以得到任何程度的高于 Poisson 分布的峰度值. 注意到 Polya-Aeppli 分布和负二项分布分别是 Poisson-ETNB 分布在 $r = 1$ 的特例和 $r \to \infty$ 的极限情况.

例 4.57 表 4-5 中的数据来源于 Hossack et al.[62], 并给出了澳大利亚机动车辆保险的索赔次数分布. 根据本节有关峰度的结论确定它的一个近似频率模型.

表 4-5 Hossack et al. 数据

索赔数	观测频率
0	565 664
1	68 714
2	5 177
3	365
4	24
5	6
6+	0

解 均值、方差和三阶中心矩分别为 0.125 461 4, 0.129 959 9, 0.140 173 7. 根据这些数字, 有

$$\frac{\mu_3-3\sigma^2+2\mu}{(\sigma^2-\mu)^2/\mu}=7.543\ 865.$$

在 Poisson–二项分布、Poisson 负二项分布、Polya-Aeppli 分布、A 类 Neyman 分布和 Poisson-ETNB 分布中, 只有最后一个比较合适. 对于这个分布, 可以由下式估计 r

$$7.543\ 865=\frac{r+2}{r+1},$$

结果为 $r=-0.847\ 185\ 1$. 在例 13.14 中将考虑更为正规的估计和模型选择方法, 但结果是相同的. □

关于复合 Poisson 分布的一个非常有用的性质是, 这个概率分布在卷积运算下是封闭的, 见以下定理.

定理 4.58 *假设 S_i 是复合 Poisson 分布, Poisson 参数为 λ_i, 次分布为 $\{q_n(i);n=0,1,2,\cdots\}$; $i=1,2,3,\cdots,k$. 并假设 $S_1,S_2,\cdots,S_k$ 是相互独立的随机变量. 则 $S=S_1+S_2+\cdots+S_k$ 为复合 Poisson 分布, Poisson 参数为 $\lambda=\lambda_1+\lambda_2+\cdots+\lambda_k$, 次分布为 $\{q_n;n=0,1,2,\cdots\}$, 其中 $q_n=[\lambda_1q_n(1)+\lambda_2q_n(2)+\cdots+\lambda_kq_n(k)]/\lambda$.*

证明 令 $Q_i(z)=\sum_{n=0}^{\infty}q_n(i)z^n$, $i=1,2,\cdots,k$. 则 S_i 的概率生成函数为 $P_{S_i}(z)=\mathrm{E}(z^{S_i})=\exp\{\lambda_i[Q_i(z)-1]\}$. 因为 S_i 相互独立, 则 S 的概率生成函数为

$$\begin{aligned}P_S(z)&=\prod_{i=1}^{k}P_{S_i}(z)=\prod_{i=1}^{k}\exp\{\lambda_i[Q_i(z)-1]\}\\&=\exp\left[\sum_{i=1}^{k}\lambda_iQ_i(z)-\sum_{i=1}^{k}\lambda_i\right]=\exp\{\lambda[Q(z)-1]\},\end{aligned}$$

其中 $\lambda=\sum_{i=1}^{k}\lambda_i$, $Q(z)=\sum_{i=1}^{k}\lambda_iQ_i(z)/\lambda$. 由生成函数的唯一性导出所求结果. □

这个结果的一个最主要的好处是易于计算. 如果关心独立复合 Poisson 随机变量的和, 则不需要计算每一个复合 Poisson 随机变量的分布. 由定理 4.58 知, 利用例 4.50 中的复合 Poisson 递推公式就足够了. 下例采用了这种方法.

例 4.59 假设 $k=2$, S_1 为复合 Poisson 分布, 变量 $\lambda_1=2$, 次分布为 $q_1(1)=0.2$, $q_2(1)=0.7, q_3(1)=0.1$. S_2(独立于 S_1) 也是 $\lambda_2=3$ 的复合 Poisson 分布, 次分布为 $q_2(2)=0.25, q_3(2)=0.6, q_4(2)=0.15$. 确定 $S=S_1+S_2$ 的分布.

解 自然有 $\lambda=\lambda_1+\lambda_2=2+3=5$. 则

$$
\begin{aligned}
q_1 &= 0.4(0.2)+0.6(0)=0.08,\\
q_2 &= 0.4(0.7)+0.6(0.25)=0.43,\\
q_3 &= 0.4(0.1)+0.6(0.6)=0.40,\\
q_4 &= 0.4(0)+0.6(0.15)=0.09.
\end{aligned}
$$

因此, S 为复合 Poisson 变量, $\lambda=5$, 次分布为 $q_1=0.08, q_2=0.43, q_3=0.4, q_4=0.09$. S 的数值分布可以由以下递推式得到

$$\Pr(S=x)=\frac{5}{x}\sum_{n=1}^{x} nq_n\Pr(S=x-n), \quad x=1,2,\cdots,$$

初始值为 $\Pr(S=0)=\mathrm{e}^{-5}$. □

在不同情况下, 负二项分布卷积的表现非常有趣. 下例表明怎样估计这些分布.

例 4.60 (负二项分布的卷积) 假设 N_i 服从参数为 (r_i,β_i) 的负二项分布, $i=1,2,\cdots,k$; 并且 $N_1,N_2,\cdots,N_k$ 相互独立. 确定 $N=N_1+N_2+\cdots+N_k$ 的分布.

解 N_i 的概率生成函数为 $P_{N_i}(z)=[1-\beta_i(z-1)]^{-r_i}$. 因此 N 的概率生成函数为 $P_N(z)=\prod\limits_{i=1}^{k}P_{N_i}(z)=\prod\limits_{i=1}^{k}[1-\beta_i(z-1)]^{-r_i}$. 如果对所有 $i=1,2,\cdots,k$, $\beta_i=\beta$, 则有 $P_N(z)=[1-\beta(z-1)]^{-(r_1+r_2+\cdots+r_k)}$, 因此 N 为参数 $(r=r_1+r_2+\cdots+r_k,\beta)$ 的负二项分布.

如果不是所有的 β_i 都相等, 证明过程如下.

由例 4.53,

$$P_{N_i}(z)=[1-\beta_i(z-1)]^{-r_i}=\mathrm{e}^{\lambda_i[Q_i(z)-1]},$$

其中 $\lambda_i=r_i\ln(1+\beta_i)$, 并且

$$Q_i(z)=1-\frac{\ln[1-\beta_i(z-1)]}{\ln(1+\beta_i)}=\sum_{n=1}^{\infty}q_n(i)z^n,$$

其中

$$q_n(i)=\frac{[\beta_i/(1+\beta_i)]^n}{n\ln(1+\beta_i)}, \quad n=1,2,\cdots.$$

但是, 定理 4.58 表明 $N = N_1 + N_2 + \cdots + N_k$ 为复合 Poisson 分布, Poisson 参数为

$$\lambda = \sum_{i=1}^{k} r_i \ln(1 + \beta_i),$$

次分布为

$$q_n = \sum_{i=1}^{k} \frac{\lambda_i}{\lambda} q_n(i) = \frac{\sum_{i=1}^{k} r_i[\beta_i/(1+\beta_i)]^n}{n \sum_{i=1}^{k} r_i \ln(1+\beta_i)}, \quad n = 1, 2, 3, \cdots .$$

N 的分布可以由以下递推式得到

$$\Pr(N = n) = \frac{\lambda}{n} \sum_{k=1}^{n} k q_k \Pr(N = n - k), \quad n = 1, 2, \cdots ,$$

以零点概率 $\Pr(N=0) = \mathrm{e}^{-\lambda} = \prod_{i=1}^{k} (1 + \beta_i)^{-r_i}$ 为初值递推, λ 和 q_n 由上面给出. □

可以看出定理 4.58 是定理 4.37 的推广, 可以用 $q_1(i) = 1$, $i = 1, 2, \cdots, k$ 代替. 类似地, 定理 4.38 的分解也可以扩展成复合 Poisson 随机变量的分解, 其分解基于次分布的支集. 更多的细节见 Panjer and Willnot[106]6.4 节, 或是 Karlin and Taylor[72] 的 16.9 节.

4.6.9 混合频率模型

也有很多的复合分布的构造方法不同于一般的复合过程. 本节考虑混合分布时, 是令模型的一个或多个参数为 "随机" 的. 本节将推广 4.6.3 节关于 gamma 混合 Poisson 分布为负二项分布的讨论.

假设模型的参数的分布位于某个样本集中, 而生成数据需要两个步骤. 首先要选择一个参数值, 在给定参数以后, 再利用该参数的分布产生观测值.

例如, 在机动车辆保险定价分析时, 因为费率的分级模式需要将个体分为 (相对) 同质的类, 用来进行分类的变量包括年龄、驾驶经验、违规记录、事故发生记录和其他变量. 因为位于同一类的个体总是会存在一定的风险残差, 混合模型将对这种差异提供一个建模的框架.

令 $P(z|\theta)$ 表示已知风险参数为 θ 的条件下, 事件发生次数的概率生成函数. 例如, 参数 θ 可以是 Poisson 分布的均值, 在这种情况下, 对风险的度量是一段时期内事件发生次数的期望值.

令 $U(\theta) = \Pr(\Theta \leqslant \theta)$ 是 Θ 的累积分布函数, Θ 是风险参数, 将其看成是一个随机变量. 则 $U(\theta)$ 表示当 Θ 被选定后 (例如样本中的一个司机), 风险参数不超过 θ 的概率. 令 $u(\theta)$ 表示 Θ 的概率函数或概率分布函数. 则

$$P(z) = \int P(z|\theta) u(\theta) \mathrm{d}\theta \text{ 或 } P(z) = \sum P(z|\theta_j) u(\theta_j) \tag{4.29}$$

表示事件次数的无条件概率生成函数 (这里选用的公式将根据 Θ 是离散或连续变量①). 相对应的概率记为

$$p_k=\int p_k(\theta)u(\theta)\mathrm{d}\theta \text{ 或 } p_k=\sum p_k(\theta_j)u(\theta_j). \tag{4.30}$$

混合分布记为 $U(\theta)$, 可以是离散或连续类型的分布或是它们的组合. **离散混合**是对离散类型分布的混合, 类似地有**连续混合**. 这种现象在 4.4.5 节的索赔量连续混合分布和 4.2.3 节的有限离散混合中有所介绍.

应该注意到被混合的分布是不可观测的, 因为观测的数据是来自混合分布的.

例 4.61 证明零点修正分布可以由二元混合分布构造.

证 假设

$$P(z)=p\cdot 1+(1-p)P_2(z).$$

这是一个零点退化分布和一个概率生成函数为 $P_2(z)$ 的分布的 (离散) 二元混合分布. 由 (4.25) 式知这是一个复合 Bernoulli 分布.

很多混合模型可以由简单分布来构造. 下面是两个例子. □

例 4.62 确定二项分布和 beta 分布的混合分布的概率函数. 这个分布被称为二项 -beta, 负二项超几何或 Polya-Eggenberger 分布.

解 beta 分布的概率分布函数为

$$u(q)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}q^{a-1}(1-q)^{b-1},\quad a>0,\quad b>0.$$

则混合分布的概率为

$$\begin{aligned}p_k&=\int_0^1\binom{m}{k}q^k(1-q)^{m-k}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}q^{a-1}(1-q)^{b-1}\mathrm{d}q\\&=\frac{\Gamma(a+b)\Gamma(m+1)\Gamma(a+k)\Gamma(b+m-k)}{\Gamma(a)\Gamma(b)\Gamma(k+1)\Gamma(m-k+1)\Gamma(a+b+m)}\\&=\frac{\binom{-a}{k}\binom{-b}{m-k}}{\binom{-a-b}{m}},\quad k=0,1,2,\cdots.\end{aligned}$$

□

例 4.63 确定参数 $p=(1+\beta)^{-1}$ 的混合负二项分布的概率函数, 其中 p 为 beta 分布. 混合分布被称为广义 Waring 分布.

① 更一般地, 我们写成 $P(z)=\int P(z|\theta)\mathrm{d}U(\theta)$, 它包含了 Θ 部分连续和部分离散的情况.

解 在例 4.62 中, 有

$$\begin{aligned} p_k &= \frac{\Gamma(r+k)}{\Gamma(r)\Gamma(k+1)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 p^{a+r-1}(1-p)^{b+k-1} \mathrm{d}p \\ &= \frac{\Gamma(r+k)}{\Gamma(r)\Gamma(k+1)} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+r)\Gamma(b+k)}{\Gamma(a+r+b+k)}, \quad k=0,1,2,\cdots. \end{aligned}$$

当 $b=1$ 时, 称为 Waring 分布. 当 $r=b=1$ 时, 为 Yule 分布. □

4.6.10 混合 Poisson

在 (4.30) 式中令 $p_k(\theta)$ 服从 Poisson 分布, 这样产生的一类分布具备很好的性质, 这种混合 Poisson 的一个简单例子是二元混合.

例 4.64 *假设可以将驾驶员分为"好驾驶员"和"差驾驶员", 每一类的索赔数都服从 Poisson 分布. 确定这个模型的概率函数, 并用其拟和例 12.56 中的数据. 这个模型和数据集来源于 Ttöbliger [130].*

解 由 (4.30) 式概率函数为

$$p_k = p\frac{\mathrm{e}^{-\lambda_1}\lambda_1^k}{k!} + (1-p)\frac{\mathrm{e}^{-\lambda_2}\lambda_2^k}{k!}.$$

由 Ttöbliger 计算的最大似然估计①为 $\hat{p}=0.94, \hat{\lambda}_1=0.11, \hat{\lambda}_2=0.7$. 这说明大约 94% 的驾驶员是"好的", 风险为每年有 $\lambda_1=0.11$ 次期望事故, 约 6% 的驾驶员是"差的", 风险为每年有 $\lambda_2=0.7$ 次期望事故. 注意我们不能够通过数据集来判断哪些驾驶员是"差驾驶员". □

这个例子说明了有限混合的两个重点. 首先, 模型可能过于简单, 而实际上很可能是一些连续风险水平的集合而不是两个风险水平. 第二点是有限混合模型有很多参数需要估计, 上例中的简单二元 Poisson 混合分布就有三个参数. 将混合的分布个数增加到 r 将在 r 元混合分布中增加 $r-1$ 个参数. 因此, 连续混合模型更受欢迎.

混合 Poisson 分布类拥有一些有趣的性质, 这里作进一步讨论.

令 $P(z)$ 表示混合 Poisson 分布的概率生成函数, 混合分布 $U(\theta)$ 任意. 则 (连续情况的公式), 引入尺度参数 λ 后, 有

$$\begin{aligned} P(z) &= \int \mathrm{e}^{\lambda\theta(z-1)} u(\theta)\mathrm{d}\theta = \int \left[\mathrm{e}^{\lambda(z-1)}\right]^{\theta} u(\theta)\mathrm{d}\theta \\ &= E\left\{[\mathrm{e}^{\lambda(z-1)}]^{\theta}\right\} = M_{\Theta}[\lambda(z-1)], \end{aligned} \tag{4.31}$$

其中 $M_{\Theta}(z)$ 为混合分布的矩母函数.

① 最大似然估计在 12.2 节中讨论.

因此, $P'(z) = \lambda M'_\Theta[\lambda(z-1)]$, 当 $z=1$ 时有 $\mathrm{E}(N) = \lambda\mathrm{E}(\Theta)$, 其中 N 是混合 Poisson 分布. 同样有 $P''(z) = \lambda^2 M''_\Theta[\lambda(z-1)]$, 推出 $\mathrm{E}[N(N-1)] = \lambda^2\mathrm{E}(\Theta^2)$, 则有

$$\begin{aligned}\mathrm{Var}(N) &= \mathrm{E}[N(N-1)] + \mathrm{E}(N) - [\mathrm{E}(N)]^2 = \lambda^2\mathrm{E}(\Theta^2) + \mathrm{E}(N) - \lambda^2[\mathrm{E}(\Theta)]^2 \\ &= \lambda^2\mathrm{Var}(\Theta) + \mathrm{E}(N) > \mathrm{E}(N).\end{aligned}$$

因此, 对于混合 Poisson 分布来说方差总是大于均值.

Douglas[29] 证明了对于任一混合 Poisson 分布, 其混合后的分布是唯一的. 这说明两个不同的混合分布不能产生同一个混合 Poisson 分布. 这使我们在某些情况下可以确定混合分布.

混合 Poisson 分布和复合 Poisson 分布之间也存在密切的关系.

定义 4.65 *称分布为是无限可分的, 若对于所有 $n = 1, 2, 3, \cdots$, 特征函数 $\varphi(z)$ 可以表示为*

$$\varphi(z) = [\varphi_n(z)]^n,$$

其中 $\varphi_n(z)$ 是某个随机变量的特征函数.

换言之, 其特征函数的 $1/n$ 次幂依然是特征函数. 特征函数的定义如下.

定义 4.66 *随机变量 X 的特征函数为*

$$\varphi_X(z) = \mathrm{E}(\mathrm{e}^{\mathrm{i}z}X) = \mathrm{E}(\cos zX + \mathrm{i}\sin zX),$$

其中 $\mathrm{i} = \sqrt{-1}$.

在定义 4.65 中, 也可以用矩生成函数或概率生成函数或其他转换形式替代特征函数. 实际上, 对于给定的随机变量, 如果满足某个转换形式, 它一定满足所有其他的形式. 之所以选择特征函数是因为所有分布的特征函数都存在, 而对于某些尾部较厚的分布其矩生成函数不一定存在. 因为之前的很多结果涉及概率生成函数, 因此概率生成函数与特征函数的关系就显得尤为重要.

定理 4.67 *如果随机变量 X 的概率生成函数存在, 则 $P_X(z) = \varphi(-\mathrm{i}\ln z), \varphi_X(z) = P(\mathrm{e}^{\mathrm{i}z})$.*

证明
$$P_X(z) = \mathrm{E}(z^X) = \mathrm{E}(\mathrm{e}^{X\ln z}) = \mathrm{E}[\mathrm{e}^{-\mathrm{i}(\mathrm{i}\ln z)\mathrm{X}}] = \varphi_X(-\mathrm{i}\ln z),$$

并且

$$\varphi_X(z) = \mathrm{E}(\mathrm{e}^{\mathrm{i}zX}) = \mathrm{E}[(\mathrm{e}^{\mathrm{i}z})^X] = P_X(\mathrm{e}^{\mathrm{i}z}).$$ □

以下分布是无限可分的：正态、gamma、Poisson 和负二项. 二项分布不是无限可分的, 因为其概率生成函数的指数 m 只能取整数值. 将 m 除以 $1, 2, 3, \cdots$ 必将得到非整数值. 事实上, 所有支集有限的分布 (概率值为正数的范围) 都不是无限可分的. 下面是一个重要的结果.

定理 4.68 假设 $P(z)$ 为混合 Poisson 分布的概率生成函数, 混合分布是无限可分的. 则 $P(z)$ 也是某个复合 Poisson 分布的概率生成函数, 并且概率生成函数可以表示为

$$P(z) = \mathrm{e}^{\lambda[P_2(z)-1]},$$

其中 $P_2(z)$ 是一个概率生成函数. 如果规定 $P_2(0) = 0$, 则 $P_2(z)$ 唯一.

可以参看 Feller[35] 第 12 章的一个证明. 任何一个无限可分的混合分布, 对应的混合 Poisson 分布可以等价的描述成一个复合 Poisson 分布. 对于某些分布, 这种分布在进行数值计算时有明显的优势, 因为当次分布确定时, 利用公式 (4.22) 可以得到概率的估计. 对于多数情况确定次分布很容易实现. 另一个优点是, 因为有两种表现分布的形式, 所以, 无需在应用时给出特定的解释. 相反地, 当某个模型适用时, 也并不意味着它是复合还是混合的结果. 例如, 索赔服从负二项分布, 并不能说明个体分布服从 Poisson 分布同时 Poisson 参数服从 gamma 分布.

例 4.69 利用上面的结果和 (4.31) 式证明 gamma 混合 Poisson 变量服从负二项分布.

解 如果混合变量是 gamma 分布, 则它的矩生成函数为 (由例 3.15 的推导, β 代替 $1/\theta$)

$$M_\Theta(t) = \left(\frac{\beta}{\beta - t}\right)^\alpha, \quad \beta > 0, \quad \alpha > 0, \quad t < \beta.$$

很明显这个分布是无限可分的, 因为 $[M_\Theta(t)]^{1/n}$ 是参数为 $(\alpha/n, \beta)$ 的 gamma 分布的矩生成函数. 则该混合 Poisson 分布的概率生成函数为

$$P(z) = \left[\frac{\beta}{\beta - \lambda(z-1)}\right]^\alpha = \left[1 - \frac{\lambda}{\beta}(z-1)\right]^{-\alpha},$$

这也是负二项分布概率生成函数的形式, 其中负二项分布的参数为 $(\alpha, \lambda/\beta)$. □

例 4.53 证明了次分布为 logarithmic 分布的复合 Poisson 分布是负二项分布. 因此这种情况下定理成立. 不难看出, 如果 $u(\theta)$ 是一个概率生成函数为 $P_\Theta(z)$ 的离散随机变量的概率函数, 则混合 Poisson 分布的概率生成函数为 $P_\Theta(\mathrm{e}^{\lambda(z-1)})$, 因此是次分布为 Poisson 分布的复合 Poisson 分布.

例 4.70 证明 A 类 Neyman 分布可以由混合分布构造.

证明 如果 (4.31) 式的混合分布的概率生成函数可表示为

$$P_\Theta(z) = \mathrm{e}^{\mu(z-1)},$$

则混合 Poisson 分布的概率生成函数为

$$P(z) = \exp\{\mu[\mathrm{e}^{\lambda(z-1)} - 1]\},$$

这代表次分布为 Poisson 分布的复合 Poisson 分布, 即 A 类 Neyman 分布. □

Holgate[60] 得到了一个更为有趣的结果, 如果被混合的分布是绝对连续和单峰的, 则混合 Poisson 分布的结果也是单峰的. 对离散函数的混合可能产生多峰性, 例如, A 类 Neyman 分布可以有不止一个峰值. 读者可以尝试计算各种两参数组合的结果.

本书中的多数连续分布都存在一个尺度参数. 这说明分布的尺度变化不改变分布的形式, 而只改变尺度参数的值. 对于混合 Poisson 分布, 概率生成函数如 (4.31) 式, λ 的任何变化等价于被混合分布尺度参数的变化, 因此当被混合分布存在尺度参数时, 为简单起见可以令 $\lambda = 1$.

例 4.71 证明逆高斯混合 Poisson 分布是参数为 $r = -0.5$ 的 Poisson-ETNB 分布.

证明 逆高斯分布见附录 A, 其概率分布函数为

$$f(x) = \left(\frac{\theta}{2\pi x^3}\right)^{1/2} \exp\left[-\frac{\theta}{2x}\left(\frac{x-\mu}{\mu}\right)^2\right], \quad x > 0,$$

为方便计算, 表示为

$$f(x) = \frac{\mu}{(2\pi\beta x^3)^{1/2}} \exp\left[-\frac{(x-\mu)^2}{2\beta x}\right], \quad x > 0,$$

其中 $\beta = \mu^2/\theta$. 则该分布的矩生成函数为 (见习题 3.24)

$$M(t) = \exp\left\{-\frac{\mu}{\beta}[(1-2\beta t)^{1/2} - 1]\right\}.$$

因此逆高斯分布是无限可分的, ($[M(t)]^{1/n}$ 是参数为 μ/n 的逆高斯分布的矩生成函数). 在 (4.31) 式中令 $\lambda = 1$, 则混合分布的概率生成函数为

$$P(z) = \exp\left(-\frac{\mu}{\beta}\{[1+2\beta(1-z)]^{1/2} - 1\}\right).$$

令

$$\lambda = \frac{\mu}{\beta}[(1+2\beta)^{1/2} - 1],$$

并且

$$P_2(z) = \frac{[1-2\beta(z-1)]^{1/2} - (1+2\beta)^{1/2}}{1-(1+2\beta)^{1/2}},$$

可以看出

$$P(z) = \exp\{\lambda[P_2(z) - 1]\},$$

其中 $P_2(z)$ 是参数为 $r = -1/2$ 的广义截断负二项分布的概率生成函数.

因此 Poisson–逆高斯分布是次分布为 ETNB($r=-1/2$) 的复合 Poisson 分布. 混合 Poisson 分布和复合 Poisson 分布的关系列在表 4-6 中. □

表 4-6 混合 Poisson 分布和复合 Poisson 分布

名　字	复合次分布	混合分布
负二项分布	对数分布	gamma
Neyman-A	Poisson	Poisson
Poisson- 逆高斯分布	ETNB($r=-0.5$)	逆高斯分布

本章主要考虑易于计算的分布. 尽管有很多其他的离散分布也适用, 我们相信这里的讨论足以用来构造能够解决大多数问题的分布类.

4.6.11 频率计算中风险暴露的作用

假设当前的业务组合包含 n 个个体, 每一个体都可能发生索赔. 令 N_j 表示第 j 个个体的索赔数. 则 $N=N_1+N_2+\cdots+N_n$. 进一步假设 N_j 独立同分布, 则

$$P_N(z)=[P_{N_1}(z)]^n.$$

假设该组合将扩充到 n^* 个, 频率为 N^*, 则

$$P_{N*}(z)=[P_{N_1}(z)]^{n^*}=[P_N(z)]^{n^*/n}.$$

因此, 如果 N 是无限可分的, N^* 的分布与 N 的形式相同, 但是参数变了.

例 4.72 研究显示, 某 300 个雇员的群体中, 工伤的索赔人数服从参数为 $\beta=0.3$、$r=10$ 的负二项分布. 确定 500 个雇员群体的索赔频率分布.

解 N^* 的概率生成函数为

$$\begin{aligned}P_{N^*}(z)&=[P_N(z)]^{500/300}=\{[1-0.3(z-1)]^{-10}\}^{500/300}\\&=[1-0.3(z-1)]^{-16.67},\end{aligned}$$

这是一个参数为 $(\beta=0.3, r=16.67)$ 的负二项分布. □

对于 $(a,b,0)$ 分布类, 除了二项分布以外的分布都具有这样的性质. 对于 $(a,b,1)$ 分布类, 都没有这样的性质. 对于复合分布, 主分布必须是无限可分的. 特别地, 复合 Poisson 分布和复合负二项分布 (包括几何分布) 在增加风险暴露的情况下仍然保持分布形式不变. 之前, 我们给出了一些零点修正分布的用处, 如果可以预期对风险暴露的调整, 即使拟合的不是很好也最好选择一个复合模型. 应注意复合模型可以在零点有很高的概率.

4.6.12 离散分布总结

我们介绍了简单的 $(a, b, 0)$ 分布类, 并推广到 $(a, b, 1)$ 分布类, 然后利用复合和混合方法构造更大的分布类. 可以用简单的递推方法计算这些分布的概率. 本节主要考虑不同分布之间的相互联系, 类似于 4.5.2 节. 一些特别的关系见表 4-7.

表 4-7 离散分布的相互关系

分布	为以下分布的特例	为以下分布的极限
Poisson	ZM Poisson	负二项
		Poisson–负二项
		Poisson–逆高斯
		Polya-Aeppli[a]
		Neyman-A[b]
ZT Poisson	ZM Poisson	ZT 负二项
ZM Poisson		ZM 负二项
几何	负二项	几何–Poisson
	ZM 几何	
ZT 几何	ZT 负二项	
ZM 几何	ZM 负二项	
对数		ZT 负二项
ZM 对数		ZM 负二项
二项	ZM 二项	
负二项	ZM 负二项	
	Poisson-ETNB	
Poisson–逆高斯	Poisson-ETNB	
Polya-Aeppli	Poisson-ETNB	
Neyman-A		Poisson-ETNB

a. 也称为 Poisson–几何.

b. 也称为 Poisson-Poisson.

由以前的讨论知 $(a, b, 0)$ 分布类是 $(a, b, 1)$ 分布类的特殊情况, 零点截断分布是零点修正分布的特殊情况. 最好通过概率生成函数来观测极限情况, 如前面证明了 Poisson 分布是负二项分布的极限情况.

我们没有列出主分布是两参数模型的复合分布, 如主分布是负二项分布或 Poisson–逆高斯分布. 这是因为这些分布通常本身就是复合 Poisson 的结果, 因此, 已经考虑了更为一般的情况. 这里收集的分布在形状上非常丰富, 然而也还会有很多其他的分布. 更多的分布讨论见 Johnson, Kots and Kemp[69], Douglas[29], Panjer and Willmot[106].

习题

4.40 考虑习题 12.96 和习题 12.98 中的数据. 给出与表 4-2 类似的计算. 对于每个数据集, 从

$(a,b,0)$ 分布类中确定一个最合适的模型.

4.41 对于以下的分布分别计算 $\Pr(N=0)$, $\Pr(N=1)$, $\Pr(N=2)$.

(a) Poisson($\lambda=4$). **(b)** 几何 ($\beta=4$).

(c) 负二项 ($r=2,\beta=2$). **(d)** 二项 ($m=8,q=0.5$).

(e) Logarithmic($\beta=4$). **(f)** ETNB($r=-0.5,\beta=4$).

(g) Poisson–逆高斯 ($\lambda=2,\beta=4$).

(h) 零点修正几何 ($p_0^M=0.5,\beta=4$).

(i) Poisson-Poisson(A 类 Neyman)($\lambda_{\text{主分布}}=4,\lambda_{\text{次分布}}=1$).

(j) Poisson-ETNB($\lambda=4,r=2,\beta=0.5$).

(k) Poisson–零点修正几何分布 ($\lambda=8,p_0^M=0.5,r=2,\beta=0.5$).

4.42 离散变量的**矩生成函数**定义为

$$M_N(z)=\mathrm{E}(\mathrm{e}^{zN})=\sum_{k=0}^{\infty}p_k\mathrm{e}^{zk}.$$

证明：$P_N(z)=M_N(\ln z)$. 利用 $\mathrm{E}(N^k)=M_N^{(k)}(0)$ 证明 $P'(1)=\mathrm{E}(N)$ 和 $P''(1)=\mathrm{E}[N(N-1)]$.

4.43 根据 Poisson 分布、负二项分布和二项分布的参数域来确定它们作为 $(a,b,0)$ 分布类的 a 和 b 的可能取值. 因为这些分布构成了全部 $(a,b,0)$ 分布类, 其他 a 和 b 的取值不会得到一个合理的概率分布. 证明：$a=-1$ 和 $b=1.5$(不在可能的取值范围内) 不会产生一个合理的分布.

4.44 证明参数为 $\beta>0$, $r>-1(r\neq 0)$ 的负二项分布, 对于任何 p_1, 由 (4.15) 式得到的 p_k 都为正值且有 $\sum\limits_{k=1}^{\infty}p_k<\infty$.

4.45 对于零点截断负二项分布, 证明当 $r\to 0$ 时的概率函数如 (4.16) 式.

4.46 证明 logarithmic 分布的概率生成函数如 (4.17) 式.

4.47 证明对于 Sibuya 分布, 即 ETNB 分布在 $-1<r<0,\beta\to\infty$ 时的特例, 它的均值不存在 (即均值所定义的和式不收敛). 因为随机变量取值非负, 因此它也不存在其他正数的高阶矩.

4.48 前面没有涉及的一个频率模型为 **Zeta 分布**. 这是一个零点截断分布, 其中 $p_k^T=k^{-(\rho+1)}/\zeta(\rho+1), k=1,2,\cdots,\rho>0$. 分母为 Zeta 函数, 可以由下式得到 $\zeta(\rho+1)=\sum\limits_{k=1}^{\infty}k^{-(\rho+1)}$. 零点修正 Zeta 分布可以用一般的方法构造, 详见 Loung and Doray[88]. 证明 Zeta 分布不是 $(a,b,1)$ 分布类.

4.49 是否所有的 $(a,b,0)$ 分布类都满足定理 4.54 的条件? 对于那些满足的, 确定可以代替定理中 θ 的参数 (或是参数的函数).

4.50 对于 $i=1,\cdots,n$, 令 S_i 为相互独立的复合 Poisson 频率分布, 其中 Poisson 参数为 λ_i, 次分布的概率生成函数为 $P_2(z)$. 注意所有 n 个变量的次分布相同. 确定 $S=S_1+\cdots+S_n$ 的分布.

4.51 证明下面三个分布是相同的: (1) 几何–几何, (2) Bernoulli- 几何, (3) 零点修正几何. 即对于这三个分布中的任意 (参数) 的一个分布, 都可以在另两个分布中找到与之具有相同的概率函数或概率生成函数的分布.

4.52 证明二项–几何分布和负二项–几何分布 (参数 r 为正整数) 等价.

4.53 证明对于任意概率生成函数, 当期望存在时, 有: $P^{(k)}(1) = \mathrm{E}[N(N-1)\cdots(N-k+1)]$, 这里的 $P^{(k)}(z)$ 表示 k 阶导数. 利用这个结果确定 (4.27) 式的三阶矩.

4.54 确定 (4.28) 式的三阶矩.

4.55 证明负二项 Poisson 复合分布和负二项 Poisson 混合分布相同.

4.56 对于 $i = 1, \cdots, n$, 令 N_i 为相互独立的参数为 λ 的混合 Poisson 分布, 令混合分布的概率生成函数为 $P_i(z)$. 证明 $N = N_1 + \cdots + N_n$ 为混合 Poisson 分布并确定其概率生成函数.

4.57 给定 $\Theta = \theta$, N 服从参数为 $\lambda\theta$ 的 Poisson 分布. 随机变量 Θ 存在一个尺度参数. 证明混合分布并不依赖于 λ 的值.

4.58 给定 $\Theta = \theta$, N 服从参数为 θ 的 Poisson 分布. 随机变量 Θ 的概率分布函数为 $u(\theta) = \alpha^2(\alpha+1)^{-1}(\theta+1)\mathrm{e}^{-\alpha\theta}, \theta > 0$. 确定混合分布的概率函数. 同时证明混合分布也是一个复合分布.

4.59 离散计数随机变量 N 的概率为 $p_n = \Pr(N = n)$, $n = 0, 1, 2, \cdots$, 令 $a_n = \Pr(N > n) = \sum\limits_{k=n+1}^{\infty} p_k$, $n = 0, 1, 2, \cdots$

(a) 证明 $\mathrm{E}(N) = \sum\limits_{n=0}^{\infty} a_n$.

(b) 证明 $A(z) = \sum\limits_{n=0}^{\infty} a_n z^n$ 与 $P(z) = \sum\limits_{n=0}^{\infty} p_n z^n$ 满足关系式 $A(z) = [1 - P(z)]/(1-z)$. 给出 $z \to 1$ 时的结果.

(c) 假设 N 服从负二项分布

$$p_n = \binom{n+r-1}{n}\left(\frac{1}{1+\beta}\right)^r\left(\frac{\beta}{1+\beta}\right)^n, \quad n = 0, 1, 2, \cdots,$$

r 为一个正数, 证明

$$a_n = \beta\sum_{k=1}^{r}\binom{n+k-1}{n}\left(\frac{1}{1+\beta}\right)^k\left(\frac{\beta}{1+\beta}\right)^n, \quad n = 0, 1, 2, \cdots.$$

(d) 假设 N 服从 Sibuya 分布, 概率生成函数为 $P(z) = 1 - (1-z)^{-r}, -1 < r < 0$, 证明

$$p_n = \frac{(-r)\Gamma(n+r)}{n!\Gamma(1+r)}, \quad n = 1, 2, 3, \cdots,$$

和

$$a_n = \binom{n+r}{n}, \quad n = 0, 1, 2, \cdots.$$

(e) 设 N 服从混合 Poisson 分布

$$p_n=\int_0^{\infty}\frac{(\lambda\theta)^n\mathrm{e}^{-\lambda\theta}}{n!}\mathrm{d}U(\theta),\quad n=0,1,2,\cdots,$$

其中 $U(\theta)$ 是累积分布函数. 证明

$$a_n=\lambda\int_0^{\infty}\frac{(\lambda\theta)^n\mathrm{e}^{-\lambda\theta}}{n!}[1-U(\theta)]\mathrm{d}\theta,\quad n=0,1,2,\cdots.$$

4.60 考虑混合 Poisson 分布

$$p_n=\Pr(N=n)=\int_0^{1}\frac{(\lambda\theta)^n\mathrm{e}^{-\lambda\theta}}{n!}U'(\theta)\mathrm{d}\theta,\quad n=0,1,\cdots,$$

其中 $U(\theta)=1-(1-\theta)^k,\ 0<\theta<1,\ k=1,2,\cdots$

(a) 证明

$$p_n=k\mathrm{e}^{-\lambda}\sum_{m=0}^{\infty}\frac{\lambda^{m+n}(m+k-1)!}{m!(m+k+n)!},\quad n=0,1,\cdots.$$

(b) 利用习题 4.59 证明

$$\Pr(N>n)=\mathrm{e}^{-\lambda}\sum_{m=0}^{\infty}\frac{\lambda^{m+n+1}(m+k)!}{m!(m+k+n+1)!}.$$

(c) 当 $k=1$ 时证明

$$p_n=\frac{1-\sum_{m=0}^{n}\lambda^m\mathrm{e}^{-\lambda}/m!}{\lambda},\quad n=0,1,2,\cdots.$$

4.61 考虑混合 Poisson 分布

$$p_n=\int_0^{\infty}\frac{(\lambda\theta)^n\mathrm{e}^{-\lambda\theta}}{n!}u(\theta)\mathrm{d}\theta,\quad n=0,1,\cdots,$$

其中 $u(\theta)$ 是正稳定分布的概率分布函数 (见 Feller[36], pp.448, 583)

$$u(\theta)=\frac{1}{\pi}\sum_{k=1}^{\infty}\frac{\Gamma(k\alpha+1)}{k!}(-1)^{k-1}\theta^{-k\alpha-1}\sin(k\alpha\pi),\quad \theta>0,$$

其中 $0<\alpha<1$. Laplace 变换为 $\int_0^{\infty}\mathrm{e}^{-s\theta}u(\theta)\mathrm{d}\theta=\exp(-s^{\alpha}),s\geqslant 0$, 证明 $\{p_n;n=0,1,\cdots\}$ 是复合 Poisson 分布, 次分布为 Sibuya 分布 (这种复合 Poisson 分布通常称为离散稳定分布).

4.62 考虑混合分布为逆高斯分布倒数的混合 Poisson 分布.

(a) 利用习题 4.36 证明这个分布是负二项分布与参数为 $r=-1/2$ 的 Poisson-ETNB 分布 (即 Poisson-逆高斯分布) 的卷积.

(b) 证明 (a) 中的混合 Poisson 分布是复合 Poisson 分布, 并确定其次分布.

第5章　保险责任调整后的索赔频率和索赔量

5.1　引　　言

我们已经看到了很多涉及随机变量函数的例子, 本章将建立这些函数和保险应用之间的联系. 本章的讨论作如下假设: 所有随机变量的支集为全体非负实数或其子集. 本章及随后章节的讨论都需要区别以下的两个变量: 一个变量是对每次损失的度量, 表示为随机变量 Y^L(若损失并没有造成赔付, 它可能取值为零); 另一个变量是对每次实际赔付的赔付金额的度量, 表示为随机变量 Y^P(如果没有造成赔付, 这个变量没有定义). 当两个变量没有本质区别时 (例如, 在确定最大赔付额时就不需要区分这两者), 上标可以省略.

5.2　免　　赔

保险合同中常含有免赔条款, 当损失 x 小于或等于某个值 d 时, 保险公司不进行赔付; 当损失超过 d 时, 保险公司支付 $x-d$. 按照第 3 章的语言, 这样的免赔可以定义如下.

定义 5.1　*普通免赔方式是将随机变量按照超额损失或者左删失平移的方法进行修正 (见定义 3.6). 两种方法的区别在于免赔方式是作用于每笔损失还是作用于每笔赔付.*

此概念及其矩计算公式已经一起介绍过. 作用于每次赔付的随机变量可表示为

$$Y^P=\begin{cases}\text{未定义}, & X\leqslant d,\\ X-d, & X>d,\end{cases}$$

作用于每次损失的随机变量可表示为

$$Y^L=\begin{cases}0, & X\leqslant d,\\ X-d, & X>d.\end{cases}$$

值得注意的是, 每次赔付随机变量可以表示为 $Y^P=Y^L|Y^L>0$, 也就是说, 每次赔付变量是在损失为正的条件下每次损失变量的条件随机变量. 对于每次赔付变量

或超额损失变量, 密度函数可以表示为

$$f_{Y^P}(y)=\frac{f_X(y+d)}{S_X(d)},\ y>0, \tag{5.1}$$

如果是离散分布, 密度函数需要替换成概率函数. 其他的重要函数有

$$S_{Y^P}(y)=\frac{S_X(y+d)}{S_X(d)},$$
$$F_{Y^P}(y)=\frac{F_X(y+d)-F_X(d)}{1-F_X(d)},$$
$$h_{Y^P}(y)=\frac{f_X(y+d)}{S_X(y+d)}=h_X(y+d).$$

注意, 作为度量每次赔付的随机变量, 超额损失变量在零点没有概率.

左删失平移变量在零点有离散概率 $F_X(d)$, 它表示损失不超过 d 而使得赔付为零的概率. 对于大于零的部分, 密度函数为

$$f_{Y^L}(y)=f_X(y+d), y>0, \tag{5.2}$$

其他重要函数如下① $(y\geqslant 0)$

$$S_{Y^L}(y)=S_X(y+d),$$
$$F_{Y^L}(y)=F_X(y+d).$$

值得注意的是, 在基于赔付变量计算索赔次数时, 改变免赔额会改变赔付的频率 (而损失的频率不会改变), 这种性质会在 5.6 节中讨论.

例 5.2　对参数为 $\alpha=3$, $\theta=2\ 000$ 的 Pareto 分布, 设普通免赔额为 500, 计算类似的量.

解　利用上面的公式, 对于超额损失随机变量有

$$f_{Y^P}(y)=\frac{3(2\ 000)^3(2\ 000+y+500)^{-4}}{(2\ 000)^3(2\ 000+500)^{-3}}=\frac{3(2\ 500)^3}{(2\ 500+y)^4},$$
$$S_{Y^P}(y)=\left(\frac{2\ 500}{2\ 500+y}\right)^3,$$
$$F_{Y^P}(y)=1-\left(\frac{2\ 500}{2\ 500+y}\right)^3,$$
$$h_{Y^P}(y)=\frac{3}{2\ 500+y}.$$

① 这里没有介绍危险率函数, 因为该函数在零点没有定义, 只要设为某个有限值即可. 而对于超额损失随机变量, 危险率函数只需要进行简单的平移.

这是一个参数为 $\alpha=3, \theta=2\,500$ 的 Pareto 分布. 其左删失平移变量的分布为

$$f_{Y^L}(y)=\begin{cases}0.488, & y=0,\\ \dfrac{3(2\,000)^3}{(2\,500+y)^4}, & y>0,\end{cases}\qquad S_{Y^L}(y)=\begin{cases}0.512, & y=0,\\ \dfrac{(2\,000)^3}{(2\,500+y)^3}, & y>0,\end{cases}$$

$$F_{Y^L}(y)=\begin{cases}0.488, & y=0,\\ 1-\dfrac{(2\,000)^3}{(2\,500+y)^3}, & y>0,\end{cases}\qquad h_{Y^L}(y)=\begin{cases}\text{未定义}, & y=0,\\ \dfrac{3}{2\,500+y}, & y>0.\end{cases}$$

图 5-1 是这个密度函数的图形, 这个修正密度是通过如下步骤得到的. 对于超损随机变量, 取原密度函数大于等于 500 的部分, 然后将选取的部分整体平 (左) 移使其从零点开始, 再乘以一个常数使得曲线下方的面积仍为 1. 对于左删失平移变量也是首先取原密度函数大于等于 500 的部分, 然后在保持数值不变的情况下整体平移至原点开始, 并将剩余的概率集中在零点, 而非扩展到整个取值域. □

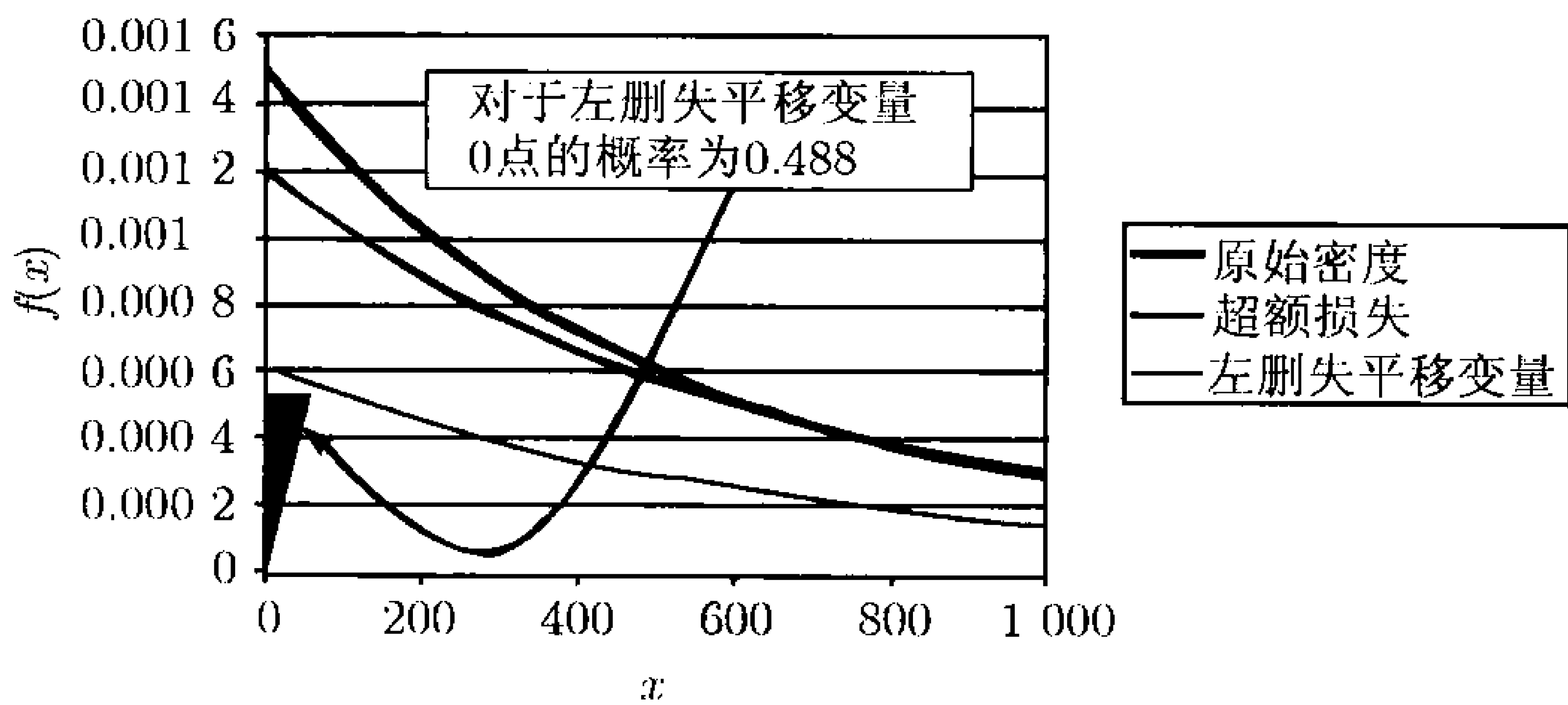


图 5-1　例 5.2 的密度函数

代替普通免赔的是特殊免赔, 它与普通免赔的不同之处在于, 当损失超过免赔额时, 所有损失 (包括免赔额以下的部分) 都将进行赔付. 例如, 在伤残保险中, 如果伤残状态在 7 天以内, 则不进行赔付; 如果伤残状态的持续时间超过 7 天, 开始进行赔偿而且赔偿金将追溯到伤残的起始日.

定义 5.3　*特殊免赔是对普通免赔的修正, 当赔付总额为正时, 特殊免赔只需在普通免赔外增加免赔额的支付.*

这里不采用左删失平移变量和超额损失的概念, 因为这种修正在保险应用中是唯一的, 我们采用每次赔付和每次损失这些术语. 这时的每次损失随机变量为

$$Y^L=\begin{cases}0, & X\leqslant d,\\ X, & X>d,\end{cases}$$

每次赔付随机变量为

$$Y^P=\begin{cases}\text{未定义}, & X\leqslant d,\\ X, & X>d,\end{cases}$$

照例, 每次赔付变量是一个条件随机变量. 前面对普通免赔得到的一些结果在这里为:

对于每次损失随机变量

$$f_{Y^L}(y)=\begin{cases}F_X(d), & y=0,\\ f_X(y), & y>d,\end{cases}\qquad S_{Y^L}(y)=\begin{cases}S_X(d), & 0\leqslant y\leqslant d,\\ S_X(y), & y>d,\end{cases}$$

$$F_{Y^L}(y)=\begin{cases}F_X(d), & 0\leqslant y\leqslant d,\\ F_X(y), & y>d,\end{cases}\qquad h_{Y^L}(y)=\begin{cases}0, & 0<y<d,\\ h_X(y), & y>d,\end{cases}$$

对于每次赔付随机变量

$$f_{Y^P}(y)=\frac{f_X(y)}{S_X(d)}, y>d, S_{Y^P}(y)=\begin{cases}1, & 0\leqslant y\leqslant d,\\ \dfrac{S_X(y)}{S_X(d)}, & y>d,\end{cases}$$

$$F_{Y^P}(y)=\begin{cases}0, & 0\leqslant y\leqslant d,\\ \dfrac{F_X(y)-F_X(d)}{1-F_X(d)}, & y>d,\end{cases}$$

$$h_{Y^P}(y)=\begin{cases}0, & 0<y<d,\\ h_X(y), & y>d.\end{cases}$$

例 5.4　对特殊免赔重新计算例 5.2.

解　使用上面的公式, 对于每次赔付变量, $y>500$ 时,

$$f_{Y^P}(y)=\frac{3(2\,000)^3(2\,000+y)^{-4}}{(2\,000)^3(2\,000+500)^{-3}}=\frac{3(2\,500)^3}{(2\,000+y)^4},$$

$$S_{Y^P}(y)=\left(\frac{2\,500}{2\,000+y}\right)^3,$$

$$F_{Y^P}(y)=1-\left(\frac{2\,500}{2\,000+y}\right)^3,$$

$$h_{Y^P}(y)=\frac{3}{2\,000+y}.$$

对于每次损失变量

$$f_{Y^L}(y)=\begin{cases}0.488, & y=0,\\ \dfrac{3(2\,000)^3}{(2\,000+y)^4}, & y>500,\end{cases}$$

$$S_{Y^L}(y)=\begin{cases}0.512, & 0\leqslant y\leqslant 500,\\ \dfrac{(2\,000)^3}{(2\,000+y)^3}, & y>500,\end{cases}$$

$$F_{Y^L}(y)=\begin{cases}0.488, & 0\leqslant y\leqslant 500,\\ 1-\dfrac{(2\ 000)^3}{(2\ 000+y)^3}, & y>500,\end{cases}$$

$$h_{Y^L}(y)=\begin{cases}0, & 0<y<500,\\ \dfrac{3}{2\ 000+y}, & y>500.\end{cases}$$

□

这两类免赔的期望成本也都可以计算.

定理 5.5 **对于普通免赔, 基于每次损失的期望成本为** $\mathrm{E}(X)-\mathrm{E}(X\wedge d)$**, 基于每次赔付的期望成本为** $[\mathrm{E}(X)-\mathrm{E}(X\wedge d)]/[1-F(d)]$**. 对于特殊免赔, 基于每次损失的期望成本为** $\mathrm{E}(X)-\mathrm{E}(X\wedge d)+d[1-F(d)]$**, 基于每次赔付的期望成本为** $[\mathrm{E}(X)-\mathrm{E}(X\wedge d)]/[1-F(d)]+d$.

证明 利用 (3.7) 式和 (3.10) 式可得, 普通免赔的每次损失随机变量的期望为 $\mathrm{E}(X)-\mathrm{E}(X\wedge d)$. 从 (5.1) 式和 (5.2) 式可看出, 要得到每次赔付随机变量的期望只需将上式除以 $1-F(d)$. 特殊免赔的赔付额比普通免赔多 d, 由此可得证明的结果. □

例 5.6 **对例 5.2 和例 5.4 中的 Pareto 分布, 计算免赔额为 500 时的上述 4 个期望值.**

解 期望值可以直接从例 5.2 和例 5.4 中求出的密度函数得到. 将定理应用到 Pareto 分布, 可以得到要求的值 (公式见附录 A). 即

$$F(500)=1-\left(\frac{2\ 000}{2\ 000+500}\right)^3=0.488,$$

$$\mathrm{E}(X\wedge 500)=\frac{2\ 000}{2}\left[1-\left(\frac{2\ 000}{2\ 000+500}\right)^2\right]=360.$$

因为 $\mathrm{E}(X)$=1 000, 所以, 对于普通免赔, 每次损失的期望成本为 1 000−360=640, 每次赔付的期望成本为 640/0.512=1 250. 对于特殊免赔, 期望成本分别为 640+500 (1−0.488)=896, 1 250+500=1 750. □

习题

5.1 对如下分布 (见 2.2 节的模型 4) 进行例 5.2 中的计算, 设普通免赔额为 5 000.

$$F_4(x)=\begin{cases}0, & x<0,\\ 1-0.3\mathrm{e}^{-0.000\ 01x}, & x\geqslant 0.\end{cases}$$

5.2 按照特殊免赔重新计算习题 5.1.

5.3 对习题 5.1 中的模型, 设免赔额为 5 000, 重新计算例 5.6.

5.4* 风险 1 服从参数为 $\alpha>2$ 和 θ 的 Pareto 分布, 风险 2 服从参数为 0.8α 和 θ 的 Pareto 分布. 每个风险均为独立的保单承保, 每张保单都含有普通免赔 k. 对风险 1 计算每次

损失的期望成本, 计算风险 2 与风险 1 每次损失的期望成本之比, 计算 k 趋于无穷时这个比值的极限.

5.5* 在不采用免赔条款时, 损失分布服从表 5-1. 设原来每次损失的普通免赔额为 10 000, 免赔额增加后使超过新免赔额的损失数目为超过原免赔额数目的一半. 计算免赔额增加后每次赔付的期望成本改变的百分比.

表 5-1　习题 5.5 的数据

x	$F(x)$	$\mathrm{E}(X\wedge x)$
10 000	0.60	6 000
15 000	0.70	7 700
22 500	0.80	9 500
32 500	0.90	11 000
∞	1.00	20 000

5.3 损失缩减率以及通货膨胀对普通免赔的影响

在评估免赔条款的影响时, 损失缩减率是一个很有价值的比率.

定义 5.7 *损失缩减率是指含有普通免赔条款后, 期望赔付的减少量与无免赔条款时的期望赔付之比.*

有很多对承保范围进行修正的方式都可以降低期望赔付, 损失缩减率专门用于描述免赔额对期望赔付的影响. 无免赔时, 期望赔付为 $\mathrm{E}(X)$; 有免赔时, 由定理 5.5 知期望赔付为 $\mathrm{E}(X)-\mathrm{E}(X\wedge d)$, 所以损失缩减率为

$$\frac{\mathrm{E}(X)-[\mathrm{E}(X)-\mathrm{E}(X\wedge d)]}{\mathrm{E}(X)}=\frac{\mathrm{E}(X\wedge d)}{\mathrm{E}(X)},$$

这里假设 $\mathrm{E}(X)$ 存在.

例 5.8 *对参数为 $\alpha=3$ 和 $\theta=2\ 000$ 的 Pareto 分布, 设普通免赔为 500, 计算损失缩减率.*

解 根据例 5.6 的计算结果, 可以马上得到损失缩减率为 360/1 000=0.36. 所以在免赔额为 500 的普通免赔条款下损失将平均降低 36%. □

通货膨胀通常会增加承保的成本. 特别地, 当应用免赔条款时, 通货膨胀的影响可能会被放大. 首先, 很多低于免赔额的损失事件在通货膨胀后会引起赔付, 另外, 因为通货膨胀调整后免赔额仍保持原值, 所以通货膨胀的相对影响也会被放大. 例如, 某事件原来产生的损失为 600 元, 如果免赔额为 500 元, 则赔付为 100 元. 当存在 10%的通货膨胀时, 事件本身的损失将提高到 660 元, 因此赔付额提高到 160 元, 最终对保险公司来讲, 因通货膨胀因素造成的损失增加比例为 60%(远远高于通货膨胀本身 10%的水平).

定理 5.9 对于含有普通免赔条款 (免赔额为 d) 的保单, 在固定的通货膨胀水平 $1+r$ 下, 每次损失的期望成本可表示为

$$(1+r)\{\mathrm{E}(X)-\mathrm{E}[X\wedge d/(1+r)]\}.$$

条件为 $F[d/(1+r)]<1$. 而每次赔付的期望成本由上式除以 $1-F[d/(1+r)]$ 得到, 条件相同.

证明 考虑通货膨胀后, 损失随机变量由 $Y=(1+r)X$ 表示, 根据定理 4.19 知, $f_Y(y)=f_X[y/(1+r)]/(1+r)$, $F_Y(y)=F_X[y/(1+r)]$, 利用 (3.8) 式, 有

$$\begin{aligned}
\mathrm{E}(Y\wedge d)&=\int_0^d yf_Y(y)\mathrm{d}y+d[1-F_Y(d)]\\
&=\int_0^d \frac{yf_X[y/(1+r)]}{1+r}\mathrm{d}y+d\left[1-F_X\left(\frac{d}{1+r}\right)\right]\\
&=\int_0^{d/(1+r)}(1+r)xf_X(x)\mathrm{d}x+d\left[1-F_X\left(\frac{d}{1+r}\right)\right]\\
&=(1+r)\left\{\int_0^{d/(1+r)}xf_X(x)\mathrm{d}x+\frac{d}{1+r}\left[1-F_X\left(\frac{d}{1+r}\right)\right]\right\}\\
&=(1+r)\mathrm{E}\left(X\wedge\frac{d}{1+r}\right),
\end{aligned}$$

其中的第 3 行进行了变量替换 $x=y/(1+r)$. 又因 $\mathrm{E}(Y)=(1+r)\mathrm{E}(X)$, 即完成了定理第一部分的证明. 而每次赔付变量的结果可通过 Y 和 X 的分布函数间的关系得到. □

例 5.10 已知损失服从参数为 $\alpha=3$ 和 $\theta=2\ 000$ 的 Pareto 分布, 设普通免赔额为 500, 计算 10%通货膨胀率的影响.

解 由例 5.6 的计算知, 每次损失变量和每次赔付变量的期望成本分别为 640 和 1 250. 对于 10%的通货膨胀有

$$\begin{aligned}
\mathrm{E}\left(X\wedge\frac{500}{1.1}\right)&=\mathrm{E}(X\wedge 454.55)\\
&=\frac{2\ 000}{2}\left[1-\left(\frac{2\ 000}{2\ 000+454.55}\right)^2\right]=336.08.
\end{aligned}$$

通货膨胀后每次损失变量的期望成本为 1.1(1 000−336.08)=730.32, 相当于增加了 14.11%. 对于每次赔付变量, 受通货膨胀影响的分布为

$$F_Y(500)=F_X(454.55)=1-\left(\frac{2\ 000}{2\ 000+454.55}\right)^3=0.459.$$

所以期望成本为 730.32/(1−0.459)=1 350, 相当于增加了 8%. □

习题

5.6 对于下面给定的分布, 设普通免赔额为 5 000, 计算损失缩减率. 这是与习题 5.1 相同的模型.

$$F_4(x)=\begin{cases}0, & x<0,\\ 1-0.3\mathrm{e}^{-0.000\ 01x}, & x\geqslant 0.\end{cases}$$

5.7 对于习题 5.6 中的分布, 设普通免赔额为 5 000, 计算 10%通货膨胀的影响.

5.8* 损失服从参数为 $\mu=7,\sigma=2$ 的 lognormal 分布, 假设免赔额为 2 000, 每年平均有 10 笔损失发生, 计算损失缩减率. 如果存在 20%的通货膨胀, 而免赔额始终为 2 000, 计算期望赔付额.

5.9* 损失服从参数为 $\alpha=2,\theta=k$ 的 Pareto 分布, 假设普通免赔额为 $2k$, 对以下两种情况计算损失缩减率: 无通货膨胀; 100%的通货膨胀.

5.10* 损失服从期望为 1 000 的指数分布, 假设免赔额为 500, 若使损失缩减率加倍, 试计算免赔额的最低增加量.

5.11* 随机变量 X 服从表 5-2 中的数据, 对每次损失变量有 15 000 元的免赔额条款, 没有承保限额. 对每次赔付变量 X 计算期望成本. 假设通货膨胀率为 50%, 试重新计算以上问题 (免赔额 15 000 元保持不变).

表 5-2　习题 5.11 的数据

x	$F(x)$	$\mathrm{E}(X\wedge x)$
10 000	0.60	6 000
15 000	0.70	7 700
22 500	0.80	9 500
∞	1.00	20 000

5.12* 损失服从参数为 $\mu=6.907\ 8,\sigma=1.517\ 4$ 的 lognormal 分布, 计算免赔额分别为 10 000 和 1 000 时的损失缩减率. 如果损失整体增加 10%, 计算损失额超过 1 000 的损失数的上升比例.

5.13* 损失的均值为 2 000, 免赔额为 1 000 时损失缩减率为 0.3, 损失额超过 1 000 的概率为 0.4, 在已知损失不高于 1 000 的条件下计算平均损失.

5.4 保 单 限 额

与免赔条款相对应的是保单限额. 在合同中较典型的保单限额条款为: 当损失低于 u 时, 保险公司赔付全部损失; 当损失超过 u 时, 保险公司只赔付 u. 限额相当于产生了一个右删失的随机变量. 这个变量为混合分布, 分布函数和密度函数分别为 (Y 是应用限额条款后的随机变量)

$$F_Y(y)=\begin{cases}F_X(y), & y<u,\\ 1, & y\geqslant u,\end{cases}$$

和

$$f_Y(y)=\begin{cases} f_X(y), & y<u, \\ 1-F_X(u), & y=u. \end{cases}$$

通货膨胀的影响可按照下面的计算进行.

定理 5.11 保单限额为 u, 通货膨胀为 $1+r$, 则期望成本调整为 $(1+r)\mathrm{E}[X\wedge u/(1+r)]$.

证明 期望成本为 $\mathrm{E}(Y\wedge u)$. 根据定理 5.9 的证明可以得到本定理的结论. □

保单限额无需区分每次赔付变量和每次损失变量的概念, 因为能够造成保险公司实际赔付的损失在应用限额条款前需要赔付, 在应用保单限额后仍然会产生赔付, 而且实际的效果相同.

例 5.12 已知损失服从参数为 $\alpha=3$ 和 $\theta=2\ 000$ 的 Pareto 分布, 设保单限额为 3 000, 计算施加限额后每次损失变量的期望成本, 以及期望成本的缩减比例. 如果考虑 10%的通货膨胀, 重新计算以上问题.

解 对于 Pareto 分布, 期望成本为

$$\mathrm{E}(X\wedge 3\ 000)=\frac{2\ 000}{2}\left[1-\left(\frac{2\ 000}{2\ 000+3\ 000}\right)^2\right]=840,$$

缩减比例为 (1 000−840)/1 000=0.16. 通货膨胀调整后的期望成本为

$$1.1\mathrm{E}(X\wedge 3\ 000/1.1)=1.1\frac{2\ 000}{2}\left[1-\left(\frac{2\ 000}{2\ 000+3\ 000/1.1}\right)^2\right]=903.11,$$

这时保单限额的缩减比例为 (1 100−903.11)/1 100=0.179. 注意, 通货膨胀调整后的期望成本增加了 7.51%, 低于通货膨胀率. 与免赔条款相反, 限额条款将缓解通货膨胀的影响, 而不会加剧.

图 5-2 显示了右删失随机变量的密度函数形状, 从 0 到 3 000 与原 Pareto 分布吻合, 超过 3 000 的概率为 $\Pr(X>3\ 000)=(2\ 000/5\ 000)^3=0.064$, 这个概率集中在 3 000 点. □

保单限额和普通免赔在概念上是相通的. 不管是哪一个方式, 当从保险公司角度看相当于从投保人角度采取了另一个方式. 例如, 如果保单条款中含有 500 的免赔, 相对投保人来说, 对每次损失的自我支付为在 500 处右删失的随机变量. 如果保单含有 3 000 的限额条款, 则投保人自身的支付就是一个左截断平移随机变量(类似一个普通免赔). 与特殊免赔方式相对应的是对任意损失的右截断方式的赔付(见习题 3.12). 这种承保责任保单很难卖出. (如果损失超过 u, 则保险公司不进行任何赔付, 你是否会购买这种保险呢？)

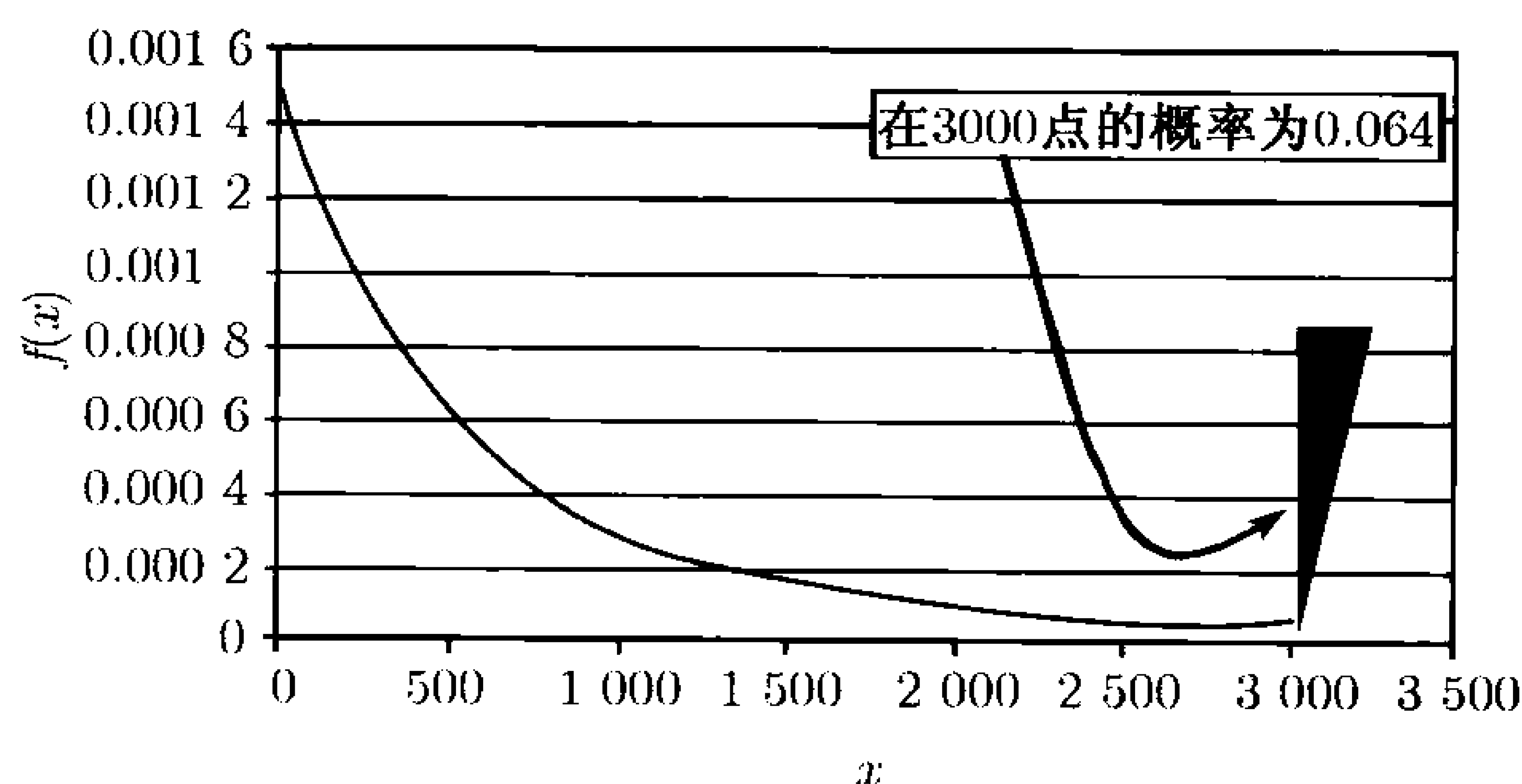


图 5-2 例 5.12 的密度函数

习题

5.14 对以下分布考虑 150 000 的保单限额, 计算 10%通货膨胀的影响. 这是与习题 5.1 和习题 5.6 相同的分布.

$$F_4(x)=\begin{cases}0, & x<0,\\ 1-0.3\mathrm{e}^{-000\ 01x}, & x\geqslant 0.\end{cases}$$

5.15* 设 X 服从参数为 $\alpha=2,\theta=100$ 的 Pareto 分布, 计算平均剩余寿命函数 $e(d)$ 的值域, d 取值于全体正实数. 令 $Y=1.1X$, 计算比值 $e_Y(d)/e_X(d)$ 的值域, d 取值于全体正实数. 最后, 令 Z 为 X 在 500 处的右删失变量 (也就是将限额 500 应用于 X), 计算 $e_Z(d)$ 的值域, d 的取值区间为 0 到 500.

5.5 分保、免赔和限额

最后一种常用的对承保责任的修正是分保方式. 在这种方式下, 保险公司按照损失的一定比例 (例如 α) 进行赔付, 投保人支付剩余的部分. 如果分保方式是对赔付责任的唯一修正, 它将使得损失变量 X 变为赔付变量 $Y=\alpha X$, 这里已经考虑了乘数的影响. 如果将本章介绍的 4 种赔付方式 (普通免赔、限额、分保和通货膨胀) 全部采用, 可得到如下表示的每次损失变量:

$$Y^L=\begin{cases}0, & X<\dfrac{d}{1+r},\\ \alpha[(1+r)X-d], & \dfrac{d}{1+r}\leqslant X<\dfrac{u}{1+r},\\ \alpha(u-d), & X\geqslant\dfrac{u}{1+r}.\end{cases}$$

在这个定义中, 各个量是按照一个特殊的顺序发生作用的. 特别地, 分保在最后一步进行. 在上面介绍的合同中, 保单限额为 $\alpha(u-d)$, 是最大的支付总额, 当损失

超过 u 时不会有额外的受益, 称 u 为**最大承保损失**. 在这里每次赔付变量 Y^P, 当 $X < d/(1+r)$ 时没有定义.

综合上面的结果可以不加证明的得到如下的定理.

定理 5.13　对于每次损失变量, 有

$$\mathrm{E}(Y^L) = \alpha(1+r)\left[\mathrm{E}\left(X \wedge \frac{u}{1+r}\right) - \mathrm{E}\left(X \wedge \frac{d}{1+r}\right)\right].$$

每次赔付变量的期望值为

$$\mathrm{E}(Y^P) = \frac{\mathrm{E}(Y^L)}{1 - F_X\left(\dfrac{d}{1+r}\right)}.$$

要得到高阶矩将更加困难, 下面的定理给出二阶矩的计算公式, 减去均值的平方可得到方差.

定理 5.14　对每次损失变量, 有

$$\begin{aligned}\mathrm{E}[(Y^L)^2] = \alpha^2(1+r)^2\{&\mathrm{E}[(X \wedge u^*)^2] - \mathrm{E}[(X \wedge d^*)^2]\\ &-2d^*\mathrm{E}(X \wedge u^*) + 2d^*\mathrm{E}(X \wedge d^*)\},\end{aligned}$$

其中 $u^* = u/(1+r)$ 和 $d^* = d/(1+r)$. 上述表达式除以 $1 - F_X(d^*)$ 即可得到每次赔付变量的二阶矩.

证明　根据 Y^L 的定义, 有

$$\begin{aligned}\mathrm{E}[(Y^L)^2] &= \int_{d^*}^{u^*} \alpha^2[(1+r)x - d]^2 f(x)\mathrm{d}x + \int_{u^*}^{\infty} \alpha^2(u-d)^2 f(x)\mathrm{d}x,\\
\frac{\mathrm{E}[(Y^L)^2]}{\alpha^2} &= (1+r)^2\left[\int_0^{u^*} x^2 f(x)\mathrm{d}x - \int_0^{d^*} x^2 f(x)\mathrm{d}x\right]\\
&\quad -2(1+r)d\left[\int_0^{u^*} x f(x)\mathrm{d}x - \int_0^{d^*} x f(x)\mathrm{d}x\right]\\
&\quad +d^2[F(u^*) - F(d^*)] + (u-d)^2[1 - F(u^*)]\\
&= (1+r)^2\{\mathrm{E}[(X \wedge u^*)^2] - u^{*2}[1 - F(u^*)]\\
&\quad -\mathrm{E}[(X \wedge d^*)^2] + d^{*2}[1 - F(d^*)]\}\\
&\quad -2(1+r)^2 d^*\{\mathrm{E}(X \wedge u^*) - u^*[1 - F(u^*)]\\
&\quad -\mathrm{E}(X \wedge d^*) + d^*[1 - F(d^*)]\}\\
&\quad +(1+r)^2 d^{*2}[F(u^*) - F(d^*)]\\
&\quad +(1+r)^2(u^* - d^*)^2[1 - F(u^*)],\\
\frac{\mathrm{E}[(Y^L)^2]}{[\alpha(1+r)]^2} &= \mathrm{E}[(X \wedge u^*)^2] - \mathrm{E}[(X \wedge d^*)^2] - 2d^*[\mathrm{E}(X \wedge u^*) - \mathrm{E}(X \wedge d^*)].\end{aligned}$$

□

例 5.15 已知参数为 $\alpha=3,\theta=2\ 000$ 的 Pareto 分布，现有免赔 500 和限额 2 500 的保单，计算每次损失变量的均值和标准差．注意，最大承保损失为 $u=3\ 000$.

解 根据前面的例子可知，$\mathrm{E}(X\wedge 500)=360$, $\mathrm{E}(X\wedge 3\ 000)=840$，限额变换的二阶矩为

$$\begin{aligned}
\mathrm{E}[(X\wedge u)^2] &= \int_0^u x^2\frac{3(2\ 000)^3}{(x+2\ 000)^4}\mathrm{d}x+u^2\left(\frac{2\ 000}{u+2\ 000}\right)^3\\
&=3(2\ 000)^3\int_{2\ 000}^{u+2\ 000}(y-2\ 000)^2y^{-4}\mathrm{d}y+u^2\left(\frac{2\ 000}{u+2\ 000}\right)^3\\
&=3(2\ 000)^3\left(-y^{-1}+2\ 000y^{-2}-\frac{2\ 000^2}{3}y^{-3}\Big|_{2\ 000}^{u+2\ 000}\right)\\
&\quad+u^2\left(\frac{2\ 000}{u+2\ 000}\right)^3\\
&=3(2\ 000)^3\left[-\frac{1}{u+2\ 000}+\frac{2\ 000}{(u+2\ 000)^2}-\frac{2\ 000^2}{3(u+2\ 000)^3}\right]\\
&\quad+3(2\ 000)^3\left[\frac{1}{2\ 000}-\frac{2\ 000}{2\ 000^2}+\frac{2\ 000^2}{3(2\ 000)^3}\right]+u^2\left(\frac{2\ 000}{u+2\ 000}\right)^3\\
&=(2\ 000)^2-\left(\frac{2\ 000}{u+2\ 000}\right)^3(2u+2\ 000)(u+2\ 000).
\end{aligned}$$

则 $\mathrm{E}[(X\wedge 500)^2]=160\ 000$, $\mathrm{E}[(X\wedge 3\ 000)^2]=1\ 440\ 000$. 所以，

$$\mathrm{E}(Y)=840-360=480,$$

$$\mathrm{E}(Y^2)=1\ 440\ 000-160\ 000-2(500)(840)+2(500)(360)=800\ 000,$$

方差为 $800\ 000-480^2=569\ 600$，标准差为 754.72. □

习题

5.16* 已知 $e(0)=25, S(x)=1-x/\omega, 0\leqslant x\leqslant\omega$, Y^P 是 d=10 时的超额损失变量，计算 Y^P 的方差.

5.17* 损失比 (R) 定义为总损失 (L) 除以已赚保费 (P). 如果公司的损失比 R 小于 0.7，代理人就可获得奖励 (B)，而且 $B=P(0.7-R)/3$，否则奖金为零. 令 $P=500\ 000$，L 服从参数为 $\alpha=3,\theta=600\ 000$ 的 Pareto 分布，计算奖金的期望值.

5.18* 已知某年的损失分布满足 $\mathrm{E}(X\wedge d)=-0.025d^2+1.475d-2.25$, d=10,11,12, $\cdots$, 26. 下一年的损失将均匀增加 10%，现有保险责任在免赔额为 11 和最大赔付额为 11 的条件下全额赔偿. 计算下一年的平均赔付额与当年平均赔付额的比例.

5.19* 已知某损失服从均值为 1 000 的指数分布，保险公司的实际赔付额为每次索赔中超过免赔额 100 的部分. 计算保险公司的每次索赔量的方差，考虑 0 点的赔付概率.

5.20* 某健康计划的总索赔服从参数为 $\alpha=2,\theta=500$ 的 Pareto 分布. 该计划向医生提供如下奖励：如果总索赔低于 500，则将 500 与索赔之差的 50%作为对医生的奖励. 这

一奖励机制将改变索赔分布，使其服从参数为 $\alpha = 2$, $\theta = K$ 的 Pareto 分布，它使得新的期望索赔加上期望奖金等于奖励机制实行前的期望索赔. 试计算 K 的值.

5.21* 已知在 a 年总期望损失为 10 000 000, 已知个体损失在 a 年服从参数为 $\alpha = 2, \theta = 2\,000$ 的 Pareto 分布. 再保公司支付每个个体损失中超过 3 000 的部分, 并且再保保费为所承保的损失的期望的 110%. 在 b 年, 损失量受到 5%通货膨胀的影响, 但损失频率没有改变. 计算 b 年保费与 a 年保费的比.

5.22* 已知损失服从 0 到 50 000 的均匀分布, 对每次损失有 5 000 的免赔和 20 000 的限额 (这意味着最大承保损失为 25 000). 在已知赔付发生的条件下, 计算期望赔付额.

5.23* 损失服从参数为 $\mu = 10, \sigma = 1$ 的对数正态分布, 如果损失低于 50 000, 不进行赔付, 如果损失在 50 000 和 100 000 之间, 赔付全部损失, 如果损失超过 100 000, 只赔付保单限额 100 000. 计算每次损失变量的期望成本.

5.6　免赔对索赔频率的影响

当采取免赔 (普通或特别) 方式时, 赔付频率的分布会随之改变, 很重要的是分析这种保单修正与赔付频率变化的影响. 当使用免赔额或增加免赔额时, 赔付频率会减少; 当免赔额降低时, 会有更多的赔付发生.

我们可以量化这个过程, 假设承保范围的修正不会影响损失本身的过程以及投保者个体的状况. 实际不然, 例如, 如果某些人购买了含 250 免赔额的机动车辆财产损失保险, 那么相对于购买全额承保的个体, 前者发生交通事故的次数可能要少一些. 类似地, 雇主们也发现, 如果对受雇的最初几年的员工减少伤残受益的支付, 永久伤残的比例就会下降.

首先, 设 X_j 表示第 j 个个体的损失量, 这里没有承保责任的任何修正, N^L 表示损失发生的数目. 现在考虑对承保责任的修正, v 是损失导致赔付的概率, 例如, 如果免赔额为 d, 则 $v = \Pr(X > d)$. 然后定义指示型随机变量 I_j, 如果第 j 次损失引起赔付, $I_j = 1$, 否则 $I_j = 0$. I_j 服从参数为 v 的 Bernoulli 分布, I_j 的 pgf 为 $P_{I_j}(z) = 1 - v + vz$. 则 $N^P = I_1 + \cdots + I_{N^L}$ 表示总赔付数. 如果 $I_1, I_2, \cdots$ 相互独立, 并且和 N^L 独立, 则 N^P 是 N^L 主分布和 Bernoulli 次分布的混合分布. 所以

$$P_{N^P}(z) = P_{N^L}[P_{I_j}(z)] = P_{N^L}[1 + v(z-1)].$$

在特殊情况下, 可以假设 N^L 的分布依赖于某个参数 θ, 且有

$$P_{N^L}(z) = P_{N^L}(z;\theta) = B[\theta(z-1)],$$

其中 $B(z)$ 是独立于 θ 的函数 (如定理 4.54), 那么

$$P_{N^P}(z) = B[\theta(1 - v + vz - 1)] = B[v\theta(z-1)] = P_{N^L}(z; v\theta).$$

这意味着 N^L 和 N^P 来自于相同的参数族, 只是参数 θ 需要改变.

例 5.16 用负二项分布证明上面的结果. 现有参数为 $r=2,\beta=3$ 的负二项分布以及 250 的免赔, 试分析免赔条款的影响. 这里假设损失服从参数为 $\alpha=3,\theta=1\ 000$ 的 Pareto 分布.

解 负二项分布的 pgf 为 $P_{N^L}(z)=[1-\beta(z-1)]^{-r}$, 这里 β 起着上述结果中 θ 的作用, $B(z)=(1-z)^{-r}$, 那么 N^P 一定服从参数为 $r^*=r,\beta^*=v\beta$ 的负二项分布. 对于题中所述的特殊情况,

$$v=1-F(250)=\left(\frac{1\ 000}{1\ 000+250}\right)^3=0.512,$$

所以 $r^*=2,\beta^*=3(0.512)=1.536.$ □

这一结果可以推广到零点修正和零点截断的分布. 假设 N^L 依赖于参数 θ 和 α, 所以

$$P_{N^L}(z)=P_{N^L}(z;\theta,\alpha)=\alpha+(1-\alpha)\frac{B[\theta(z-1)]-B(-\theta)}{1-B(-\theta)}. \tag{5.3}$$

注意, $\alpha=P_{N^L}(0)=\Pr(N^L=0)$, 所以它是零点的修正概率. 如果 $B[\theta(z-1)]$ 本身也是 pgf, 则 (5.3) 式给出的 pgf 与零点修正的分布一致. 然而, 为了使 (5.3) 式中的 $P_{N^L}(z)$ 成为 pgf, 并不一定要求 $B[\theta(z-1)]$ 为 pgf. 特别地, $B(z)=1+\ln(1-z)$ 产生了零点修正 (ZM) 的 logarithmic 分布, 尽管没有概率分布以这个 $B(z)$ 作为 pgf. 类似地, $B(z)=(1-z)^{-r}, -1<r<0$ 产生了 ETNB 分布. 经过简单的代数推导就可揭示出对于 (5.3) 式有,

$$P_{N^P}(z)=P_{N^L}(z;v\theta,\alpha^*),$$

其中 $\alpha^*=\Pr(N^P=0)=P_{N^P}(0)=P_{N^L}(1-v;\theta,\alpha)$. 当应用免赔条款时, α 的值会增加, 因为无赔付的可能性将增加. 特别地, 如果 N^L 是零点截断的, 则 N^P 是零点修正的.

例 5.17 重新计算上一个例子, 令频率分布为零点修正的参数为 $r=2,\beta=3$ 的负二项分布, $p_0^M=0.4$.

解 pgf为

$$P_{N^L}(z)=p_0^M+(1-p_0^M)\frac{[1-\beta(z-1)]^{-r}-(1+\beta)^{-r}}{1-(1+\beta)^{-r}}.$$

那么 $\alpha=p_0^M$, $B(z)=(1-z)^{-r}$, 进而有 $r^*=r$, $\beta^*=v\beta$, 并且

$$\begin{aligned}\alpha^*=p_0^{M*}&=p_0^M+(1-p_0^M)\frac{(1+v\beta)^{-r}-(1+\beta)^{-r}}{1-(1+\beta)^{-r}}\\&=\frac{p_0^M-(1+\beta)^{-r}+(1+v\beta)^{-r}-p_0^M(1+v\beta)^{-r}}{1-(1+\beta)^{-r}}.\end{aligned}$$

对于这个给定的特殊分布, 新参数为 $r^* = 2$, $\beta^* = 3(0.512) = 1.536$, 和

$$p_0^{M*} = \frac{0.4 - 4^{-2} + 2.536^{-2} - 0.4(2.536)^{-2}}{1 - 4^{-2}} = 0.459\ 5. \qquad \square$$

在应用中, 我们可能会由 N^P 求 N^L 的分布. 例如, 我们搜集到的数据可能是应用免赔条款后的赔付数, 从这些数据可以估计出 N^P 的参数. 如果想了解去掉免赔条款后的赔付分布, 正如前面所讨论的,

$$P_{N^L}(z) = P_{N^P}(1 - v^{-1} + zv^{-1}).$$

这意味着只要将原来的 v 替换成 $1/v$ 就可以得到所求方程. 可是, 这时得到 N^L 的 pgf 可能是无效的, 可能因为模型的某一假设是无效的 (例如, 假设免赔的改变不会改变索赔行为).

例 5.18　假设对含有 250 免赔额的保单的赔付服从零点修正的负二项分布, $r^* = 2$, $\beta^* = 1.536$, $p_0^{M*} = 0.459\ 5$. 损失服从参数为 $\alpha = 3$, $\theta = 1\ 000$ 的 Pareto 分布. 计算免赔条款去除后赔付数的分布. 如果假设 $p_0^{M*} = 0.002$, 重新计算该问题.

解　在第一种情况下 v=1/0.512=1.953 125, 所以 r=2, β=1.953 125(1.536)=3. 而且

$$P_0^{M*} = \frac{0.459\ 5 - 2.53\ 6^{-2} + 4^{-2} - 0.459\ 5(4)^{-2}}{1 - 2.536^{-2}} = 0.4,$$

第二种情况,

$$P_0^{M*} = \frac{0.002 - 2.536^{-2} + 4^{-2} - 0.002(4)^{-2}}{1 - 2.536^{-2}} = -0.107\ 9,$$

这不是一个可行的概率值. □

所有 $(a,b,0)$ 和 $(a,b,1)$ 类的分布都满足这一节的条件, 表 5-3 显示了由 N^L 转到 N^P 时参数是如何变化的. 如果 N^L 服从复合分布, 则可以表示为 $P_{N^L}(z) = P_1[P_2(z)]$, 故

$$P_{N^P}(z) = P_{N^L}[1 + v(z-1)] = P_1\{P_2[1 + v(z-1)]\}.$$

所以 N^P 也服从复合分布, 次分布按下面所显示的进行修正. 如果次分布为 $(a,b,0)$ 类分布, 可以按表 5-3 修正. 下面的例子显示了次分布为 $(a,b,1)$ 类分布时的调整方法.

例 5.19　假设 N^L 服从 Poisson-ETNB, $\lambda = 5, \beta = 0.3, r = 4$. 如果 $v = 0.5$, 计算 N^P 的分布.

解 根据上面的讨论知, N^P 服从 $\lambda^* = 5$ 的复合 Poisson 分布, 而次分布为零点修正的负二项分布 (根据表 5-3)$\beta^* = 0.5(0.3) = 0.15$,

$$p_0^{M*} = \frac{0 - 1.3^{-4} + 1.15^{-4} - 0(1.15)^{-4}}{1 - 1.3^{-4}} = 0.341\ 03,$$

$r^* = 4$. 这些已经足够了, 除非我们已经有了使用 ETNB 作为次分布的习惯. 根据定理 4.54, 一个带有零点修正的次分布的复合 Poisson 分布等价于一个带有零点截断的次分布的复合 Poisson 分布, Poisson 参数必须变为 $(1 - p_0^{M*})\lambda^*$. 所以, N^P 服从 Poisson-ETNB 分布, $\lambda^* = (1 - 0.341\ 03)5 = 3.294\ 85$, $\beta^* = 0.15$ 和 $r^* = 4$. □

表 5-3 频率的调整

N^L	N^P 的参数
Poisson	$\lambda^* = v\lambda$
ZM Poisson	$p_0^{M*} = \dfrac{p_0^M - e^{-\lambda} + e^{-v\lambda} - p_0^M e^{-v\lambda}}{1 - e^{-\lambda}}, \lambda^* = v\lambda$
二项	$q^* = vq$
ZM 二项	$p_0^{M*} = \dfrac{p_0^M - (1-q)^m + (1-vq)^m - p_0^M(1-vq)^m}{1-(1-q)^m}$
	$q^* = vq$
负二项	$\beta^* = v\beta,\ r^* = r$
ZM 负二项	$p_0^{M*} = \dfrac{p_0^M - (1+\beta)^{-r} + (1+v\beta)^{-r} - p_0^M(1+v\beta)^{-r}}{1-(1+\beta)^{-r}}$
负二项	$\beta^* = v\beta,\ r^* = r$
ZM 对数	$p_0^{M*} = 1 - (1 - p_0^M)\ln(1+v\beta)/\ln(1+\beta)$
	$\beta^* = v\beta$

这一结果可以进一步推广到增加或减少免赔额的情景. 令 N^d 表示免赔为 d 时的索赔频率, N^{d*} 表示免赔为 d^* 时的索赔频率, $v = [1 - F_X(d^*)]/[1 - F_X(d)]$, 则表 5-3 可用来表示由 N^d 转化到 N^{d*} 时参数的改变. 只要 $d^* > d$, 则 $v < 1$, 方程导出了 N^{d*} 的一个合理分布, 它也包含了本节开始时使用的 $d = 0$ 的特殊情况. 如果 $d^* < d$, 则 $v > 1$, 此时不保证得到的结果是一个合理的分布. 它包括前面提过的 $d^* = 0$ 的情况 (消除免赔条款).

最后需要注意的是, 保单限额对频率分布没有影响, 无论是使用、去除或改变限额都不会改变赔付的数目.

习题

5.24 某团体人身保险保单含有意外死亡的附加条款, 如果正常死亡, 赔偿金为 10 000, 如果意外死亡, 赔偿金为 20 000. 近似认为被保险人年龄相同, 所以假设他们有相同的索赔概率是合理的. 设不发生索赔的概率为 0.97, 正常死亡的概率为 0.01, 意外死亡的概率为 0.02. 现有再保险公司提供超额再保险, 对每一个意外死亡的个体支付 10 000.

(a) 可按照如下过程建立索赔模型，频率部分服从 Bernoulli 分布 (事件是发生索赔或不发生索赔)，损失部分有两点取值 (在给定索赔发生的条件下，概率与两种索赔水平相联系). 具体写出频率变量和损失变量的概率分布.

(b) 假设再保险公司仍然希望保持相同的频率分布，计算修正的损失分布，使其能够反映再保险公司的赔付情况.

(c) 计算再保险公司的频率变量和损失变量的概率分布，这里的损失变量分布是指再保险公司实际进行赔付的条件下的损失.

5.25 个体损失服从参数为 $\alpha = 2,\ \theta = 1\ 000$ 的 Pareto 分布，免赔额为 500 时，赔付数的频率分布服从参数为 $\lambda = 3,\ \beta = 2$ 的 Poisson-inverse Gaussian 分布. 如果免赔额增加至 1 000，计算赔付数的分布. 对于新的免赔额，计算每次赔付变量的损失分布的 pdf.

5.26 损失服从参数为 $\alpha = 2,\ \theta = 1\ 000$ 的 Pareto 分布，免赔额为 500 时，频率分布为零点截断的 logarithmic 分布，参数为 $\beta = 4$. 当免赔额减少到 0 时，确定赔付数的模型.

5.27 假设损失的数目 N^L 服从 Sibuya 分布 (见习题 4.47 和习题 4.59)，它的 pgf 为 $P_{N^L}(z) = 1 - (1-z)^{-r}, -1 < r < 0$. 证明：赔付数服从零点修整的 Sibuya 分布.

第6章　总损失模型

6.1　引　　言

本章的主要目的是推导总损失模型 (aggregate loss model). 该模型表示在给定的时期内给定的保单组合的所有赔案的总赔付额. 一般有两种方法对赔付额进行加总, 以获得一段时期内的总赔付额.

一种方法是记录那些发生索赔的赔付额并进行加总, 这样总损失随机变量 S 表示对个体赔付额 $(X_1, X_2, \cdots, X_N)$ 的随机的赔案数 N 之和, 即

$$S = X_1 + X_2 + \cdots + X_N, \quad N = 0, 1, 2, \cdots, \tag{6.1}$$

且当 $N = 0$ 时, $S = 0$.

定义 6.1　聚合风险模型(collective risk model) 具有形如 (6.1) 的表达式, 若无特别说明, X_j 均为独立同分布 (i.i.d.) 的随机变量. 用更为正式的语言叙述的独立性假设为:

(1) 在 $N = n$ 的条件下, 随机变量 $X_1, X_2, \cdots, X_n$ 为 i.i.d. 的随机变量;

(2) 在 $N = n$ 的条件下, 随机变量 $X_1, X_2, \cdots, X_n$ 的共同分布不依赖 n;

(3) N 的分布不以任何的方式依赖于 $X_1, X_2, \cdots$ 的取值.

下面给出另一个模型的描述.

定义 6.2　个体风险模型(individual risk model) 将总损失表示为某个固定的 n 项的和式 $S = X_1 + \cdots + X_n$, 其中 n 为保险合同数. 这 n 份合同各自的损失额为 $(X_1, X_2, \cdots, X_n)$, 其中 X_j 独立但并不一定同分布. X_j 的分布通常在零点有概率质量, 表示无损失或赔付发生的概率.

个体风险模型用于加总固定保单数的或保单组合的损失或赔付额. 它可用在对一组包含 n 个雇员的团体寿险或健康险的损失建模中, 每个雇员会有不同的保障水平 (寿险的受益水平是工资的倍数), 也会有不同的损失概率 (不同的年龄和健康状况).

在 X_j 同分布的特殊情形下, 个体风险模型就成了聚合风险模型的特例, N 的分布变为全部概率都集中在 $N = n$ 上的退化分布, 即 $\Pr(N = n) = 1$.

(6.1) 式中 S 的分布由 N 的分布和 X_j 的分布得到, 利用这种方法可以分别对索赔频率和索赔量建模, 用索赔频率和索赔量分布的信息来得到 S 的信息. 另一种

方法是直接收集 S 的信息 (如一段时间内的月度总损失) 并利用第 4 章里的模型对 S 建模.

对 N 和 X_j 的分布分别建模有一些显而易见的好处.

(1) 赔案数的期望值随保单数的变化而变化. 在基于往年的数据预测未来的赔案数时, 需要将业务量的增长考虑在内.

(2) 一般的宏观经济通货膨胀和新增赔案的通货膨胀效应都会反映在被保险个体的损失额以及保险公司的赔付额上. 若保单的免赔额和限额不随通货膨胀调整, 那么采用总损失的个体模型时上述的通货膨胀效应常常被掩盖.

(3) 改变个体保单的免赔额和限额的影响更容易研究, 可通过改变索赔量分布的性质来进行.

(4) 更容易理解因免赔额的变化而引起的索赔频率变化.

(5) 可以将免赔额和保单限额不同的保单的数据混合在一起, 以得到假定的损失额分布, 这对于整合来自不同年份且保单条款不同的数据时非常有意义.

(6) 对以下各个损失部分的建模是具有内在一致性的：投保人的未承保损失、保险人的理赔成本和再保险人的理赔成本. 直接保险人研究再保转移损失的影响时, 这种性质非常有意义.

(7) N 和 X_j 的分布形状共同决定了 S 的分布形状. 了解相对的形状在调整保单细节时非常有用. 例如, 若索赔量分布具有比索赔频率更厚的尾部, 则总赔付额或总损失的尾部形状将由索赔量的分布来决定, 而对索赔频率分布的选择并不敏感.

总之, 分别考虑索赔频率和损失程度, 可以构造出更准确、更灵活的模型.

在对 S 构造模型 (6.1) 式时, 若 N 表示实际发生的损失数, 那么 X_j 可以表示为：(i) 投保人的损失; (ii) 保险人的赔付额; (iii) 再保险人的赔付额; (iv) 投保人承担的免赔额 (自保额). 在每种情况下, S 的含义各不相同, 索赔额的分布也可以根据不同的含义做出相应的调整.

由于在本章和随后第 7 章和第 8 章里会多次使用随机变量 N, $X_1, X_2, \cdots$ 和 S, 我们希望在使用时要特别谨慎. 称 N 为**索赔次数随机变量**(claim count random variable), 有时也称**赔案数**或简称**赔案**, 称 N 的分布为**索赔次数分布**(claim count distribution). 另一个常用术语为**索赔频率分布**. 也称 X_j 为**个体**或者**单次损失随机变量**(individual or single-loss random variable), 当指代对象清楚时, 修饰语**个体**或**单次**可以去掉. 在第 5 章中, 定义了损失与赔付的不同概念. 严格地说, X_j 是赔付因为它反映了真实的现金交易. 尽管如此, 习惯上仍可称之为**损失**, 而且还会继续使用. 另一个表示 X_j 的常用术语是**损失程度**. 最后, 称 S 为**总损失随机变量**(aggregate loss random variable).

例 6.3　某保险人通过估计已知被保险个体的损失所服从的概率密度函数为 $f_X(x)$,

保险人的赔付责任为个体损失超过 1 000 部分的 80%, 最大赔付额为 100 000, 并对赔付额超过 50 000 的部分进行再保. 试分别为以下各个损失量建立模型: (a) 被保险个体投保前的总损失; (b) 保险人实施再保险前的总损失; (c) 再保险人的总损失; (d) 再保险赔付后, 保险人的总损失; (e) 被保险个体的总损失.

解 (a) 被保险个体没有保险时的总损失为 $S = X_1 + X_2 + \cdots + X_N$, 这里 X_j 服从概率密度函数为 $f_X(x)$ 的分布.

(b) 再保险赔付前, 保险人的总赔付为 $S = Y_1 + Y_2 + \cdots + Y_N$, 其中

$$Y_j = \begin{cases} 0, & X_j \leqslant 1\ 000, \\ 0.80(X_j - 1\ 000), & 1\ 000 < X_j \leqslant 126\ 000, \\ 100\ 000, & X_j > 126\ 000. \end{cases}$$

(c) 再保险人的总赔付为 $S = Y_1 + Y_2 + \cdots + Y_N$, 其中

$$Y_j = \begin{cases} 0, & X_j \leqslant 63\ 500, \\ 0.80(X_j - 63\ 500), & 63\ 500 < X_j \leqslant 126\ 000, \\ 50\ 000, & X_j > 126\ 000. \end{cases}$$

(d) 再保险赔付后, 保险人的总支付为 $S = Y_1 + Y_2 + \cdots + Y_N$, 其中

$$Y_j = \begin{cases} 0, & X_j \leqslant 1\ 000, \\ 0.80(X_j - 1\ 000), & 1\ 000 < X_j \leqslant 63\ 500, \\ 50\ 000, & X_j > 63\ 500. \end{cases}$$

(e) 被保险个体的总支付, 即保险责任除外的损失部分为 $S = Y_1 + Y_2 + \cdots + Y_N$, 其中

$$Y_j = \begin{cases} X_j, & X_j \leqslant 1\ 000, \\ 800 + 0.20X_j, & 1\ 000 < X_j \leqslant 126\ 000, \\ X_j - 100\ 000, & X_j > 126\ 000. \end{cases}$$

□

习题

6.1 证明例 6.3 中投保人、保险人和再保险人的支付总和等于总损失.

6.2 模 型 选 择

在确定拟合频率和损失程度数据的分布时, 一般都会有很多可选的分布. 不过, 出于实用的考虑某些分布可能显得更加合适.

一般来说, 尺度分布族 (定义 4.2) 作为损失分布的拟合比较合适. 因为对这类分布, 货币币种 (如美元或英磅) 的选择不会影响拟合结果. 同时, 尺度分布族有利于调整通货膨胀的影响 (本质上还是币种的变化问题, 如 1994 年的美元与 1995 年的美元). 当预测来年的成本时, 只需调整参数就可以反映预期通货膨胀率的影响.

也可以类似地考虑索赔频率的分布. 假定其他条件相同, 随着保险公司某块业务的增长, 索赔数也会相应的增加. 若选择如下以 α 为参数的概率生成函数形式

$$P_N(z;\ \alpha) = Q(z)^\alpha, \tag{6.2}$$

则索赔数的期望值与 α 成比例. 当业务量增加 $100r\%$ 时, 索赔数的期望值将按照比例 $\alpha^* = (1+r)\alpha$ 增加, 这已经在 4.6.11 节讨论过了. 因为 r 可以是满足 $r > -1$ 的任何值, 所以满足 (6.2) 式的分布应当允许 α 取任何正数. 可以证明, 这样的分布是无限可分的 (定义 4.65).

另一种类似的现实考虑也同样支持频率分布应具有无限可分性. 这涉及某种关于研究时间的不变性的概念 —— 在理想的情形下, 建立的模型不应该受索赔频率研究时所选择的时间长度的影响, 无论对业务量的增长进行怎样的调整, 频率的期望都应该与时间长度成比例. 这意味着, 对 10 年数据的研究结果可以用来得到关于一个月、一年或是任何时间长度的索赔频率分布. 更进一步地, 一年期索赔频率的分布形式在调整参数后, 应该与一个月的分布形式相同. 参数 α 应该与时间长度相对应, 例如, 若在 (6.2) 式中一月期的 $\alpha = 1.7$, 那么参数 $\alpha = 20.4$ 的相同模型就应该适合于一年期的数据.

在零点有修正的分布不具有 (6.2) 式的形式. 但是, 如果常识认为需要进行零点修正, 我们也仍然需要考虑零点修正的分布. 例如, 若由于保险责任的重复性或其他原因, 某部分保单从来没有发生索赔, 那么在未来一段时间内, 将按照这个比例随机选取这类保单.

习题

6.2 若概率生成函数满足 (6.2) 式的分布, 证明其均值与 α 成比例.

6.3 附录 B 中的哪些分布对任何正数 α 都符合 (6.2) 式?

6.3 总索赔的复合模型

令 S 为 N 个满足 (6.1) 式中独立性假设的索赔额观测值 $X_1, X_2, \cdots, X_N$ 的总损失, 本章的处理方法为:

(1) 基于数据建立 N 的分布模型;

(2) 基于数据建立 X_j 的共同分布模型;

(3) 使用以上两个模型, 通过必要的计算获得 S 的分布.

前两个步骤将由本书其他章节的内容来完成, 现在假设这两个模型已经建立, 则只需要通过数值计算的工作来解答有关 S 分布的问题. 这可能涉及止损再保险合同的定价问题, 需要分析个体损失的免赔额、分保的比例和最大赔付额等带来的影响.

随机和

$$S = X_1 + X_2 + \cdots + X_N$$

的 (N 服从记数分布) 分布为

$$F_S(x) = \Pr(S \leqslant x) = \sum_{n=0}^{\infty} p_n \Pr(S \leqslant x | N = n) = \sum_{n=0}^{\infty} p_n F_X^{*n}(x), \tag{6.3}$$

其中 $F_X(x) = \Pr(X \leqslant x)$ 是 X_j 共同的分布函数, $p_n = \Pr(N = n)$. 在 (6.3) 式中, 称 $F_X^{*n}(x)$ 为 X 分布函数的 "n 重卷积". 它可以由

$$F_X^{*0}(x) = \begin{cases} 0, & x < 0, \\ 1, & x \geqslant 0, \end{cases}$$

和

$$F_X^{*k}(x) = \int_{-\infty}^{\infty} F_X^{*(k-1)}(x-y) \mathrm{d}F_X(y),\ k = 1, 2, \cdots \tag{6.4}$$

得到. 若 X 为在负值上概率为零的连续型随机变量, (6.4) 式化简为

$$F_X^{*k}(x) = \int_0^x F_X^{*(k-1)}(x-y) f_X(y) \mathrm{d}y,\ k = 2, 3, \cdots.$$

当 $k = 1$ 时, 上式简化为 $F_X^{*1}(x) = F_X(x)$. 通过微分, 得到概率密度函数

$$f_X^{*k}(x) = \int_0^x f_X^{*(k-1)}(x-y) f_X(y) \mathrm{d}y,\ k = 2, 3, \cdots.$$

对于在 $0, 1, 2, \cdots$, 点具有正概率值的离散型随机变量, (6.4) 式化简为

$$F_X^{*k}(x) = \sum_{y=0}^{x} F_X^{*(k-1)}(x-y) f_X(y),\ x = 0, 1, \cdots,\ k = 2, 3, \cdots.$$

对应的概率函数为

$$f_X^{*k}(x) = \sum_{y=0}^{x} f_X^{*(k-1)}(x-y) f_X(y),\ x = 0, 1, \cdots,\ k = 2, 3, \cdots.$$

分布 (6.3) 式也称做**复合分布**(compound distribution), 其总损失分布的概率函数为

$$f_S(x)=\sum_{n=0}^{\infty}p_n f_X^{*n}(x).$$

类似 4.6.7 节的讨论, 对固定的 n, 由 $X_1,\cdots,X_n$ 的独立性知 S 的概率生成函数为

$$\begin{aligned}P_S(z)&=\mathrm{E}[z^S]\\&=\sum_{n=0}^{\infty}\mathrm{E}[z^{X_1+X_2+\cdots+X_n}|N=n]\mathrm{Pr}(N=n)\\&=\sum_{n=0}^{\infty}\mathrm{E}\left[\prod_{j=1}^{n}z^{X_j}\right]\mathrm{Pr}(N=n)\\&=\sum_{n=0}^{\infty}\mathrm{Pr}(N=n)[P_X(z)]^n\\&=\mathrm{E}[P_X(z)^N]=P_N[P_X(z)].\end{aligned}\tag{6.5}$$

对其他的概率生成函数也存在类似的关系, 有时用特征函数更方便

$$\varphi_S(z)=\mathrm{E}(\mathrm{e}^{\mathrm{i}zS})=P_N[\varphi_X(z)],$$

因为特征函数总是存在的. Panjer and Willmot[106] 使用 Laplace 变换

$$L_S(z)=\mathrm{E}(\mathrm{e}^{-zS})=P_N[L_X(z)].$$

它对任何定义在非负值上的随机变量都是存在的. 考虑到矩母函数, 有

$$M_S(z)=P_N[M_X(z)].$$

复合分布的概率生成函数在 4.6.7 节讨论过, 其中的 "次分布" 就相当于本章中的索赔额的分布.

若出现 $P_N(z)=P_1[P_2(z)]$ 的情况, 即 N 本身就是复合分布, 则有 $P_S(z)=P_1\{P_2[P_X(z)]\}$, 公式本身的难度没有增加.

由 (6.5) 式, S 的矩可以由 N 和 X_j 的矩得到. 前三阶矩为

$$\begin{aligned}\mathrm{E}(S)&=\mu'_{S1}=\mu'_{N1}\mu'_{X1}=\mathrm{E}(N)\mathrm{E}(X),\\\mathrm{Var}(S)&=\mu_{S2}=\mu'_{N1}\mu_{X2}+\mu_{N2}(\mu'_{X1})^2,\\\mathrm{E}\{[S-E(S)]^3\}&=\mu_{S3}=\mu'_{N1}\mu_{X3}+3\mu_{N2}\mu'_{X1}\mu_{X2}+\mu_{N3}(\mu'_{X1})^3.\end{aligned}\tag{6.6}$$

这里, 第一个下标表示相应的随机变量, 第二个下标表示矩的阶数, 上标 $'$ 表示原点矩 (关于零点的矩), 若无上标 $'$ 则表示中心矩 (关于均值的矩). 也可以用这些矩本身建立近似的总损失概率分布模型, 只需匹配样本的前几阶矩即可.

例 6.4 过去 10 个月的索赔次数和个体损失的观测均值 (和标准差) 分别为 6.7(2.3) 和 179 247(52 141). 求每月总赔付的均值和方差.

解

$$\mathrm{E}(S) = 6.7(179\ 247) = 1\ 200\ 955,$$

$$\mathrm{Var}(S) = 6.7(52\ 141)^2 + (2.3)^2(179\ 247)^2 = 1.881\ 80 \times 10^{11}.$$

因此, 总赔付的均值和标准差分别为 1 200 955 和 433 797. □

例 6.5 (例 6.4 续) 利用正态分布和对数正态分布作为总赔付的近似分布, 求总赔付额超过期望值 140% 的概率, 即

$$\Pr(S > 1.40 \times 1\ 200\ 955) = \Pr(S > 1\ 681\ 337).$$

解 对正态分布, 有

$$\begin{aligned}\Pr(S > 1\ 681\ 337) &= \Pr\left(\frac{S - \mathrm{E}(S)}{\sqrt{\mathrm{Var}(S)}} > \frac{1\ 681\ 337 - 1\ 200\ 955}{433\ 797}\right)\\ &= \Pr(Z > 1.107) = 1 - \Phi(1.107) = 0.134.\end{aligned}$$

由附录 A, 对数正态分布的均值和二阶原点矩为

$$\mathrm{E}(S) = \exp\left(\mu + \frac{1}{2}\sigma^2\right), \quad \mathrm{E}(S^2) = \exp(2\mu + 2\sigma^2).$$

分别令其等于 $1.200\ 955 \times 10^6$和$1.881\ 80 \times 10^{11} + (1.200\ 95 \times 10^6)^2 = 1.630\ 47 \times 10^{12}$, 并对结果取对数, 得到下面 2 个包含 2 个未知数的等式:

$$\mu + \frac{1}{2}\sigma^2 = 13.998\ 63, \quad 2\mu + 2\sigma^2 = 28.119\ 89.$$

得到 $\mu = 13.937\ 31$ 且 $\sigma^2 = 0.122\ 636\ 1$, 所以

$$\begin{aligned}\Pr(S > 1\ 681\ 337) &= 1 - \Phi\left[\frac{\ln 1\ 681\ 337 - 13.937\ 31}{(0.122\ 636\ 1)^{0.5}}\right]\\ &= 1 - \Phi(1.135\ 913) = 0.128.\end{aligned}$$

当 $\mathrm{E}(N)$ 较大时, 正态分布提供了一种很好的近似. 特别地, 若 N 服从 Poisson、二项或负二项分布, 中心极限定理指出, 当 λ, m 或 r 分别趋向于无穷时, S 的分布趋向于正态. 在这个例子里 $\mathrm{E}(N)$ 较小, 所以 S 的分布很可能是有偏的, 尽管没有理论说明对数正态分布的优良性, 但对数正态的确提供了一个很好的近似. □

例 6.6 (团体牙医保险) 现有如下的雇员团体牙医保险计划: 保险责任包括雇员和其家庭成员, 每个已婚员工无论其家庭成员人数是多少, 保费都是相同的. 统计

数据显示每人每年通过该计划获得的口腔护理费用 (已经根据当前的美元价值进行了调整) 具有表 6-1 所示的分布 (单位是 25 美元). 保险公司将依据该统计结果进行核算.

表 6-1 例 6.6 的损失分布

x	$f_X(x)$
1	0.150
2	0.200
3	0.250
4	0.125
5	0.075
6	0.050
7	0.050
8	0.050
9	0.025
10	0.025

此外, 每年每个承保个体对应的口腔护理受益人数 (每个雇员的家庭成员人数) 服从表 6-2 所示的分布.

表 6-2 例 6.6 的频率分布

n	p_n
0	0.05
1	0.10
2	0.15
3	0.20
4	0.25
5	0.15
6	0.06
7	0.03
8	0.01

保险公司需要估算该团体每年每个已婚员工的承保支出的分布, 每个已婚员工支出的分布满足

$$f_S(x) = \sum_{n=0}^{8} p_n f_X^{*n}(x).$$

求 S 的概率函数在 525 以下各点的值, 并计算每位雇员总赔付额的均值和标准差.

解 表 6-3 给出了 S 在 525 以下的分布. 要得到 $f_S(x)$, 需将 $f_X(x)$ 卷积矩阵的每一行都乘以表格最下方一行相对应的概率并加总.

表 6-3 例 6.6 的加总概率

x	f_X^{*0}	f_X^{*1}	f_X^{*2}	f_X^{*3}	f_X^{*4}	f_X^{*5}	f_X^{*6}	f_X^{*7}	f_X^{*8}	$f_S(x)$
0	1	0	0	0	0	0	0	0	0	0.050 00
1	0	0.150	0	0	0	0	0	0	0	0.015 00
2	0	0.200	0.022 50	0	0	0	0	0	0	0.023 38
3	0	0.250	0.060 00	0.003 38	0	0	0	0	0	0.034 68
4	0	0.125	0.115 00	0.013 50	0.000 51	0	0	0	0	0.032 58
5	0	0.075	0.137 50	0.034 88	0.002 70	0.000 08	0	0	0	0.035 79
6	0	0.050	0.135 00	0.061 44	0.008 78	0.000 51	0.000 01	0	0	0.039 81
7	0	0.050	0.107 50	0.085 69	0.019 99	0.001 98	0.000 09	0.000 00	0	0.043 56
8	0	0.050	0.088 13	0.097 50	0.035 80	0.005 49	0.000 42	0.000 02	0.000 00	0.047 52
9	0	0.025	0.078 75	0.098 41	0.052 66	0.011 94	0.001 36	0.000 08	0.000 00	0.049 03
10	0	0.025	0.070 63	0.093 38	0.066 82	0.021 38	0.003 45	0.000 31	0.000 02	0.051 90
11	0	0	0.062 50	0.088 13	0.075 97	0.032 82	0.007 26	0.000 91	0.000 07	0.051 38
12	0	0	0.045 00	0.083 70	0.080 68	0.044 50	0.013 05	0.002 18	0.000 22	0.051 19
13	0	0	0.031 25	0.076 73	0.082 66	0.054 86	0.020 62	0.004 48	0.000 60	0.050 30
14	0	0	0.017 50	0.066 89	0.082 78	0.063 14	0.029 30	0.008 08	0.001 38	0.048 18
15	0	0	0.011 25	0.053 77	0.080 81	0.069 34	0.038 26	0.013 04	0.002 79	0.045 76
16	0	0	0.007 50	0.041 25	0.075 84	0.073 61	0.046 77	0.019 19	0.005 05	0.042 81
17	0	0	0.005 00	0.030 52	0.068 11	0.075 78	0.054 38	0.026 16	0.008 29	0.039 38
18	0	0	0.003 13	0.022 67	0.058 54	0.075 52	0.060 80	0.033 52	0.012 54	0.035 75
19	0	0	0.001 25	0.016 73	0.048 78	0.072 63	0.065 73	0.040 83	0.017 68	0.031 97
20	0	0	0.000 63	0.011 86	0.039 77	0.067 47	0.068 82	0.047 75	0.023 51	0.028 32
21	0	0	0	0.008 00	0.031 87	0.060 79	0.069 82	0.053 89	0.029 77	0.024 79
p_n	0.05	0.10	0.15	0.20	0.25	0.15	0.06	0.03	0.01	

读者也许可以用 (6.6) 式来验证 $f_S(x)$ 分布的前两阶矩为

$$\mathrm{E}(S) = 12.58, \quad \mathrm{Var}(S) = 58.746\,4.$$

因此该牙医保险的年费用均值为 $12.58 \times 25 = 314.50$ 美元, 标准差为 191.615 5 美元. (为什么不能由表 6-3 计算?) □

对一段时间内的总损失采取免赔处理也是保险经营中常考虑的方法. 当这种损失处理作用在投保人身上时, 称作保险责任, 当作用于保险公司时, 称作再保险责任. 后者是保险公司在流年不利的情况下保护自身的常用方法 (即单笔特大赔案的情形). 更正式地, 给出如下的定义.

定义 6.7 对总损失考虑一定的免赔额的保险, 称作止损保险(stop-loss insurance), 这种保险的期望成本称做净止损保费(net stop-loss premium), 可由 $\mathrm{E}[(S-d)_+]$ 计算得到, 这里 d 为免赔额, 符号 $(\cdot)_+$ 表示如果括号内的值为正, 则取这个值, 否则取零.

对任何总损失分布, 有

$$\mathrm{E}[(S-d)_+]=\int_d^{\infty}[1-F_S(x)]\mathrm{d}x.$$

若分布是连续的, 净止损保费可以直接由定义计算得到

$$\mathrm{E}[(S-d)_+]=\int_d^{\infty}(x-d)f_S(x)\mathrm{d}x.$$

类似地, 对离散型随机变量, 有

$$\mathrm{E}[(S-d)_+]=\sum_{x>d}(x-d)f_S(x).$$

当存在某个区间其中任何点的概率均为零时, 下面的结论可以简化计算.

定理 6.8　*假设* $\Pr(a<S<b)=0$, *那么, 对任意* $a\leqslant d\leqslant b$, *有*

$$\mathrm{E}[(S-d)_+]=\frac{b-d}{b-a}\mathrm{E}[(S-a)_+]+\frac{d-a}{b-a}\mathrm{E}[(S-b)_+].$$

即, 净止损保费可以通过线性插值计算.

证明　由假设, $F_S(x)=F_S(a)$, 当 $a\leqslant x\leqslant b$, 那么

$$\begin{aligned}\mathrm{E}[(S-d)_+]&=\int_d^{\infty}[1-F_S(x)]\mathrm{d}x\\&=\int_a^{\infty}[1-F_S(x)]\mathrm{d}x-\int_a^{d}[1-F_S(x)]\mathrm{d}x\\&=\mathrm{E}[(S-a)_+]-\int_a^{d}[1-F_S(a)]\mathrm{d}x\\&=\mathrm{E}[(S-a)_+]-(d-a)[1-F_S(a)].\end{aligned}\tag{6.7}$$

在 (6.7) 式中令 $d=b$ 可得

$$\mathrm{E}[(S-b)_+]=\mathrm{E}[(S-a)_+]-(b-a)[1-F_S(a)],$$

因此, 有

$$1-F_S(a)=\frac{\mathrm{E}[(S-a)_+]-\mathrm{E}[(S-b)_+]}{b-a}.$$

将其代入公式 (6.7) 得到结论.　□

在离散情形, 只要 S 在等间距点上有概率, 则可以得到进一步的简化.

定理 6.9　*假设对某个固定的* $h>0$, *有* $\Pr(S=kh)=f_k\geqslant 0$, $k=0,1,\cdots$. *对其他所有的* x, *有* $\Pr(S=x)=0$. *那么, 给定* $d=jh$, j *为非负整数, 则有*

$$\mathrm{E}[(S-d)_+]=h\sum_{m=0}^{\infty}\{1-F_S[(m+j)h]\}.$$

证明

$$\begin{aligned}\mathrm{E}[(S-d)_+] &= \sum_{x>d}(x-d)f_S(x)\\ &= \sum_{k=j}^{\infty}(kh-jh)f_k\\ &= h\sum_{k=j}^{\infty}\sum_{m=0}^{k-j-1} f_k\\ &= h\sum_{m=0}^{\infty}\sum_{k=m+j+1}^{\infty} f_k\\ &= h\sum_{m=0}^{\infty}\{1-F_S[(m+j)h]\}.\end{aligned}$$ □

在等间距的数值点有概率的离散情形下, 有一个简单的递归式成立.

推论 6.10 *在定理 6.9 的条件下, 有*

$$\mathrm{E}\{[S-(j+1)h]_+\} = \mathrm{E}[(S-jh)_+] - h[1-F_S(jh)].$$

这个结果使用起来很方便, 因为当 $d=0$ 时, $\mathrm{E}[(S-0)_+]=\mathrm{E}(S)=\mathrm{E}(N)\mathrm{E}(X)$, 这可以直接由频率和损失额的分布得到.

例 6.11 (续例 6.6) *保险公司正在考虑对每位雇员的总损失进行免赔处理的影响. 试求免赔额为 20, 30, 50, 100 美元时, 净保费的减少量.*

解 由表 6-3, 在 0, 25, 50, 75 美元处的累积分布函数值分别为 0.05, 0.065, 0.088 38, 0.123 06. 由 E(S)=25(12.58)=314.5 有

$$\begin{aligned}\mathrm{E}[(S-25)_+] &= 314.5-25(1-0.05)=290.75,\\ \mathrm{E}[(S-50)_+] &= 290.75-25(1-0.065)=267.375,\\ \mathrm{E}[(S-75)_+] &= 267.375-25(1-0.088\,38)=244.584\,5,\\ \mathrm{E}[(S-100)_+] &= 244.584\,5-25(1-0.123\,06)=222.661.\end{aligned}$$

由定理 6.8, $\mathrm{E}[(S-30)_+]=20/25\times 290.75+5/25\times 267.375=286.07$. 与原净保费 314.5 相比, 4 种免赔额净保费的减少量分别为 23.75, 28.43, 47.125, 91.839. □

习题

6.4 利用 (6.5) 式证明 (6.6) 式的各阶矩之间的关系成立.

6.5 已知投保个体的住院医疗费用情况如下:

费用	均值	标准差
住院	1 000	500
其他	500	300

而且住院费和其他费用之间的协方差为 100 000. 现有如下的保单：住院费 100% 赔付、其他费用赔付 80%. 已知住院的病人数服从参数为 4 的 Poisson 分布, 求这种保单赔付的均值和标准差.

6.6 总索赔数模型为参数 $r = 15$ 和 $\beta = 5$ 的复合负二项分布, 索赔额服从 (0, 10) 上的均匀分布. 使用正态近似, 求使得总索赔额超过保费的概率为 0.05 的保费额.

6.7 现将机动车驾驶员分为三种类型, 每种类型的索赔数服从参数为 λ 的 Poisson 分布. 利用下面数据, 求随机任意选取的驾驶员的索赔数的方差.

类型	占总体的比例	λ
1	0.25	5
2	0.25	3
3	0.50	2

6.8 设 X_1, X_2 和 X_3 是相互独立的损失随机变量, 概率函数如表 6-4 所示. 求 $S = X_1 + X_2 + X_3$ 的概率函数.

表 6-4 例 6.8 的分布

x	$f_1(x)$	$f_2(x)$	$f_3(x)$
0	0.90	0.50	0.25
1	0.10	0.30	0.25
2	0.00	0.20	0.25
3	0.00	0.00	0.25

6.9 设 X_1, X_2 和 X_3 是相互独立的随机变量, 概率函数如表 6-5 所示. 若 $S = X_1 + X_2 + X_3$ 且 $f_S(5) = 0.06$, 求 p.

表 6-5 例 6.9 的分布

x	$f_1(x)$	$f_2(x)$	$f_3(x)$
0	p	0.6	0.25
1	$1 - p$	0.2	0.25
2	0	0.1	0.25
3	0	0.1	0.25

6.10 现有如下关于艾滋病的护理费用信息

(1) 已知没有患艾滋病时, 个体医疗护理费用的条件分布的均值为 1 000、方差为 250 000.

(2) 已知患有艾滋病的条件下, 个体医疗护理费用的条件分布的均值为 70 000、方差为 1 600 000.

(3) 由 m 个随机选取的成人构成的群体中患艾滋病的人数服从参数为 m 和 $q = 0.01$ 的二项分布.

保险公司用群体总索赔额分布的均值加上标准差的 10% 作为该群体的保费. 对由 10 个确定未患艾滋病的独立个体构成的群体, 保费为 P; 对由 10 个随机挑选的成人构成的群体, 保费为 Q, 求 P/Q.

6.11 城市规划人员请你帮助分析人们在办公室的吸烟情况, 并提供了如表 6-6 所示的一个工作日内吸烟次数的分布信息.

表 6-6 例 6.11 的数据

	男性	女性
均值	6	3
方差	64	31

在随机选取的由 N 位雇员共用的办公室中男性雇员数服从参数为 N 和 0.4 的二项分布. 求随机挑选的由 8 位雇员共用的办公室在一个工作日内吸烟总数的均值和标准差之和.

6.12 已知某团体的总索赔额为 (0, 10) 上的均匀分布. 保险人 A 提供免赔额为 6 的止损保险计划, 保费等于止损赔付的期望值; 保险人 B 的计划是保费为 7 个货币单位的团体分红保险 (保费返还), 当索赔额低于某个数值 $7k$ 时, 分红额为 $7k$ 超过索赔额的部分. 求使得两个计划的期望成本相同的 k 值.

6.13 假设某团体健康保险合约的总索赔额服从指数分布, 现要求核保人员来估计该分布的均值. 保险公司对总索赔额超过其期望值 125%的部分提供止损保险, 保费为止损赔偿量的期望值的两倍. 后来发现核保人员对索赔额期望值的估计有误差, 致使估计值为正确值的 90%, 求实际的附加保费比率.

6.14 随机损失额 X 具有如下的概率函数:

x	$f(x)$
0	0.05
1	0.06
2	0.25
3	0.22
4	0.10
5	0.05
6	0.05
7	0.05
8	0.05
9	0.12

已知 $\mathrm{E}(X) = 4$ 和 $\mathrm{E}[(X-d)_+] = 2$, 求 d.

6.15 设再保险人赔付总索赔额超过 d 的部分, 为此得到止损保费 $\mathrm{E}[(S-d)_+]$. 已知 $\mathrm{E}[(S-100)_+] = 15, \mathrm{E}[(S-120)_+] = 10$, 总索赔额大于 80 且小于或等于 120 的概率为 0, 求总索赔额小于或等于 80 的概率.

6.16 损失随机变量 X 的概率密度函数为 $f(x) = 1/100, 0 < x < 100$, 可以购买以下两种保单来减轻损失对财务造成的影响.

$$A = \begin{cases} 0, & x < 50k, \\ \dfrac{x}{k} - 50, & x \geqslant 50k, \end{cases}$$

以及

$$B = kx, \quad 0 < x < 100,$$

其中 A 和 B 是损失为 x 时的赔付额. 两种保单具有相同的净保费, 即 $\mathrm{E}(A) = \mathrm{E}(B)$, 求 k.

6.17 现有一份家庭医疗护理保单, 护理的平均时间为 440 天, 30%的家庭医疗护理在前 30 天就结束了, 且结束的时刻在 30 天内均匀分布. 保单规定在家庭护理的前 30 天每天赔付 20 美元, 此后每天赔付 100 美元. 求每次医疗护理赔付额的期望值.

6.18 某保单组合产生了 N 次索赔, 其中

n	$\Pr(N = n)$
0	0.5
1	0.4
3	0.1

个体赔付额具有如下分布

x	$f_X(x)$
1	0.9
10	0.1

个体赔付额与 N 相互独立, 计算总赔付额与赔付额期望值之比超过 3.0 的概率.

6.19 某公司出售团体旅游意外保险, 对每个在旅游途中意外身故的受益个体赔付 m 元. 团体的总保费等于团体总赔付额的期望值加上标准差. 标准保费的确定基于以下假设:

(1) 这个群体中的所有个体赔案是相互独立的;

(2) $m^2 q(1 - q) = 2\ 500$, 其中 q 为个体旅游的意外身故概率.

在由 100 个个体构成的团体中, 由于其中 3 个个体常常一同出行, 如果 1 个个体身故 3 个个体都会身故, 导致独立性假设不成立. 求这个团体的总保费和标准保费之差.

6.20 某寿险公司现有 16 000 个个体的一年期寿险组合如下表所示:

受益金额	投保人数	索赔概率
1	8 000	0.025
2	3 500	0.025
4	4 500	0.025

所有的赔案是相互独立的, 保险公司的自留额上限为每个个体 2 个单位, 购买再保险的保费为每单位 0.03. 这个保险公司的自留赔付额 S 与再保保费之和超过 1 000 的概率为 $\Pr\left[\frac{S-\mathrm{E}(S)}{\sqrt{\mathrm{Var}(S)}} > K\right]$, 利用正态近似求 K.

6.21 个体损失 Y 的概率密度函数为

$$f(y)=\begin{cases} 0.02\left(1-\dfrac{y}{100}\right), & 0<y<100, \\ 0, & \text{其他}. \end{cases}$$

赔付额 Z 为超过免赔额 10 以上部分的 80%, 求 $\mathrm{E}(Z)$.

6.22 个体损失分布服从参数 $\mu = 100, \sigma^2 = 9$ 的正态分布, 赔案数 N 的分布在表 6-7 中给出. 求总赔付额超过 100 的概率.

表 6-7 习题 6.22 的分布

n	$\Pr(N=n)$
0	0.5
1	0.2
2	0.2
3	0.1

6.23 某雇主的自保寿险计划具有以下特征:

(1) 已知赔案发生的情况下, 索赔额为 2 000 元的概率为 0.4、为 3 000 元的概率为 0.6.

(2) 赔案数的分布在表 6-8 中给出.

表 6-8 习题 6.23 的分布

n	$f(n)$
0	1/16
1	1/4
2	3/8
3	1/4
4	1/16

该雇主购买了一个总量止损再保险, 将其每年的赔付成本限制在 5 000 元以内, 该止损再保险的成本为 1 472 元. 求雇主的这个计划的年均支出的期望值, 包括止损再保险的成本.

6.24 已知个体患者住院时间小于或等于 k 天的概率为 $1-0.8^k, k=0,1,2,\cdots$, 住院补助保单从住院第 4 天到第 10 天 (即最多 7 天) 每天提供固定额度的补偿. 若提供补助的最长时间由 7 天增加到 14 天, 求每个住院者的期望赔付额增加的百分比.

6.25 总赔付额 S 的概率密度函数为 $f_S(x)=3x^{-4}, x\geqslant 1$, 相对附加系数 θ 和 λ 满足

$$\Pr[S\leqslant(1+\theta)\mathrm{E}(S)]=\Pr\left[S\leqslant \mathrm{E}(S)+\lambda\sqrt{\mathrm{Var}(S)}\right]=0.90.$$

求 θ 和 λ.

6.4 解析结果

对于大多数 N 和 $X_1, X_2, \cdots$ 的分布, 都只能得到复合分布的数值解. 本章下面几节将专门解决这一类数值计算问题.

但是, 如果选择一些特殊的分布, 则可以得到简单的解析结果, 计算的复杂程度将得到极大的简化.

例 6.12 (几何一指数复合) 设 $X_1, X_2, \cdots$ 是独立同分布序列, 服从均值为 θ 的指数分布且矩母函数为 $M_X(z) = (1-\theta z)^{-1}$. 设 N 服从参数为 β 的几何分布且概率生成函数为 $P_N(z) = [1-\beta(z-1)]^{-1}$(见附录 B). 求 S 的分布.

解 通过一些代数计算得到 S 的矩母函数为

$$\begin{aligned} M_S(z) &= P_N[M_X(z)] \\ &= \{1-\beta[(1-\theta z)^{-1}-1]\}^{-1} \\ &= \frac{1}{1+\beta}1 + \frac{\beta}{1+\beta}[1-\theta(1+\beta)z]^{-1}. \end{aligned}$$

这是一个在零点概率为 1 的退化分布和一个均值为 $\theta(1+\beta)$ 的指数分布的混合. 因此, $\Pr(S=0) = (1+\beta)^{-1}$, 且对 $x>0$, S 的概率密度函数为

$$f_S(x) = \frac{\beta}{\theta(1+\beta)^2}\exp\left[-\frac{x}{\theta(1+\beta)}\right].$$

它在 0 点有概率质量 $(1+\beta)^{-1}$, 而在正半轴上具有指数速率衰减的密度函数. 它的累积分布函数可以表示为

$$F_S(x) = 1 - \frac{\beta}{1+\beta}\exp\left[-\frac{x}{\theta(1+\beta)}\right], \quad x \geqslant 0.$$

此函数除 0 点有跳跃外连续. 这个例子会在第 8 章讨论破产理论时再介绍. □

例 6.13 (指数损失程度) 求任意损失程度为指数分布的复合模型中 S 的累积分布函数.

解 n 个独立的均值为 θ 的指数随机变量之和的矩母函数为

$$M_{X_1+X_2+\cdots+X_n}(z) = (1-\theta z)^{-n},$$

这是 gamma 分布的矩母函数, 对应的累积分布函数为

$$F_X^{*n}(x) = \Gamma\left(n; \frac{x}{\theta}\right)$$

(见附录 A). 对取整数值的 α, $\Gamma(\alpha; x)$ 值可以由下式精确计算 (推导见附录 A):

$$\Gamma(n;x) = 1 - \sum_{j=0}^{n-1}\frac{x^j \mathrm{e}^{-x}}{j!}, \quad n = 1,2,3,\cdots. \tag{6.8}$$

由 (6.3) 式, 有

$$F_S(x) = p_0 + \sum_{n=1}^{\infty} p_n \Gamma\left(n; \frac{x}{\theta}\right).$$

代入 (6.8) 式得到

$$F_S(x) = 1 - \sum_{n=1}^{\infty} p_n \sum_{j=0}^{n-1} \frac{(x/\theta)^j \mathrm{e}^{-x/\theta}}{j!}, \quad x \geqslant 0. \tag{6.9}$$

改变求和顺序得到

$$\begin{aligned} F_S(x) &= 1 - \mathrm{e}^{-x/\theta} \sum_{j=0}^{\infty} \frac{(x/\theta)^j}{j!} \sum_{n=j+1}^{\infty} p_n \\ &= 1 - \mathrm{e}^{-x/\theta} \sum_{j=0}^{\infty} \bar{P}_j \frac{(x/\theta)^j}{j!}, \quad x \geqslant 0, \end{aligned}$$

其中 $\bar{P}_j = \sum\limits_{n=j+1}^{\infty} p_n,\ j = 0, 1, \cdots$. □

例 6.13 的方法可以推广到范围更大的混合 Erlang 损失分布, 见习题 6.35.

对在所有非负整数上都有正概率的损失频率分布, 计算 (6.9) 式时只要取出第一个求和式的足够多项来计算即可. 对满足 $\Pr(N > n^*) = 0$ 的分布, 第一个和式变成有限的, 例如, 对二项分布的损失频率, (6.9) 式变为

$$F_S(x) = 1 - \sum_{n=1}^{m} \binom{m}{n} q^n (1-q)^{m-n} \sum_{j=0}^{n-1} \frac{(x/\theta)^j \mathrm{e}^{-x/\theta}}{j!}. \tag{6.10}$$

例 6.14 (负二项—指数复合) 设索赔频率分布是参数 r 为整数的负二项分布, 损失程度服从指数分布, 求它们的复合模型 S 的分布.

解 S 的矩母函数为

$$\begin{aligned} M_S(z) &= P_N[M_X(z)] \\ &= P_N[(1-\theta_z)^{-1}] \\ &= \{1 - \beta[(1-\theta_z)^{-1} - 1]\}^{-r}. \end{aligned}$$

通过代数运算, 它可以改写成

$$M_S(z) = \left(1 + \frac{\beta}{1+\beta}\{[1 - \theta(1+\beta)_z]^{-1} - 1\}\right)^r,$$

上式可表示为

$$M_S(z) = P_N^*[M_X^*(z)],$$

其中

$$P_N^*(z)=\left[1+\frac{\beta}{1+\beta}(z-1)\right]^r,$$

是参数为 r 和 $\beta/(1+\beta)$ 的二项分布的概率生成函数, $M_X^*(z)$ 是均值为 $\theta(1+\beta)$ 的指数分布的矩母函数.

这样的转化使得分布函数的计算可简化为形如 (6.10) 的有限和, 即

$$F_S(x)=1-\sum_{n=1}^{r}\binom{r}{n}\left(\frac{\beta}{1+\beta}\right)^n\left(\frac{1}{1+\beta}\right)^{r-n}$$

$$\times\sum_{j=0}^{n-1}\frac{[x\theta^{-1}(1+\beta)^{-1}]^j\mathrm{e}^{-x\theta^{-1}(1+\beta)^{-1}}}{j!}.$$ □

例 6.15 (卷积下封闭的损失分布) 若某分布族中独立同分布的随机变量之和的分布仍属于该分布族, 则称该分布族满足卷积封闭性(closed under convolution). 进一步假设, 卷积封闭分布族中 n 个独立同分布随机变量的和所产生的新的随机变量, 其分布除了一个参数变为原来的 n 倍外, 其他的参数均不变. 若损失分布具有上述性质时, 求复合模型 S 的分布.

解 由已知条件, 若每个 X_j 的概率函数是 $f_X(x;a)$, 则 $X_1+X_2+\cdots+X_n$ 的概率函数是 $f_X(x;na)$, 这意味着

$$f_S(x)=\sum_{n=1}^{\infty}p_nf_X^{*n}(x;a)=\sum_{n=1}^{\infty}p_nf_X(x;na),$$

这样就免去了计算卷积的繁琐. 具有卷积封闭性的损失分布包括 gamma 分布和逆高斯分布, 见习题 6.26. □

习题

6.26 考虑以下为关于卷积封闭性的问题.

(a) 证明 gamma 分布和逆高斯分布具有卷积封闭性, 并证明 gamma 分布具有例 6.15 中的 "进一步假设" 部分的性质.

(b) 离散分布同样可以作为损失程度的分布, 附录 B 中的哪些分布具有卷积封闭性? 如何利用这些信息来简化公式 (4.20) 式中的复合概率计算?

6.27 现有某参数为 $\beta=1,r=2$ 的复合负二项分布, 损失分布为 $\{f_X(x);x=0,1,2,\cdots\}$, 若损失分布改变为 $\{g_X(x)=f_X(x)/[1-f_X(0)];x=0,1,2,\cdots\}$, 应当如何修改参数才能使得总赔付额的分布不变?

6.28 考虑损失程度为指数分布的复合对数分布.

(a) 证明总损失的密度可以表示为

$$f_S(x)=\frac{1}{\ln(1+\beta)}\sum_{n=1}^{\infty}\frac{1}{n!}\left[\frac{\beta}{\theta(1+\beta)}\right]^n x^{n-1}\mathrm{e}^{-x/\theta}.$$

(b) 将上式化简为

$$f_S(x)=\frac{\exp\{-x/[\theta(1+\beta)]\}-\exp(-x/\theta)}{x\ln(1+\beta)}.$$

6.29 某保单的赔偿责任为：当总损失费用超过 100 元的部分位于第一个 1 000 元以内时赔偿超过 1 00 元部分的 80%, 位于第二个 1 000 元部分时赔偿超过部分的 90%, 位于其他部分时 100% 赔偿. 对于不同的 d 值, 用 $R_d=\mathrm{E}[(S-d)_+]$ 给出该保险责任的期望赔付的表达式.

6.30 已知投保的司机在一年内发生交通事故的次数服从参数 $\lambda=2$ 的 Poisson 分布, 当事故发生时, 损失超过免赔额的概率为 0.25, 事故数与损失额是独立的. 求在一年内没有发生损失超过免赔额的事故的概率.

6.31 某保险公司使用指数分布为总损失分布建模, 但均值尚不能确定, 公司认为均值应该服从 (2 000 000, 4 000 000) 上的均匀分布. 求总损失的期望值.

6.32 某团体医疗补助保单对最多 30 天的住院期提供每天 100 元的连续保险金给付, 保险金对不到一天的住院期是按比例计算的. 对 $0\leqslant t\leqslant 30$, 住院期 T(按天计) 具有如下的连续 (生存) 函数：

$$\Pr(T\geqslant t)=\begin{cases}1-0.04t, & 0\leqslant t\leqslant 10,\\ 0.95-0.035t, & 10<t\leqslant 20,\\ 0.65-0.02t, & 20<t\leqslant 30.\end{cases}$$

在一个保单期限内, 每位投保人住院一次的概率为 0.1, 超过一次的概率为 0. 求每位投保人的净保费, 忽略资金的时间价值.

6.33 假设普通医疗赔案和口腔医疗赔案的发生是独立地服从如下的复合 Poisson 分布：

索赔类型	索赔额分布	λ
普通医疗赔案	(0, 1 000) 上的均匀分布	2
口腔医疗赔案	(0, 200) 上的均匀分布	3

对一份同时涵盖了普通医疗和口腔医疗的保单, 令 X 为给定赔案的赔付额. 对任一给定的赔案, 求赔偿金 (超过 100 的部分) 的期望值 $\mathrm{E}[(X-100)_+]$.

6.34 某投保人的索赔总额分布是参数为 $m=12$ 和 $q=0.25$ 的二项分布, 若保险人愿意支付分红 D, 其数额等于保费的 80%超过索赔额的部分 (如果确实超过). 已知保费为 5, 求 $\mathrm{E}[D]$.

6.35 考虑一个损失分布, 它由有限个形状参数为整数的 gamma 分布 (这种 gamma 也叫作 Erlang 分布) 混合而成, 即

$$f_X(x)=\sum_{k=1}^{r}q_k\frac{\theta^{-k}x^{k-1}\mathrm{e}^{-x/\theta}}{(k-1)!},\quad x>0.$$

(a) 证明矩母函数可表示为

$$M_X(z)=Q\{(1-\theta_z)^{-1}\},$$

其中

$$Q(z)=\sum_{k=1}^{r}q_kz^k$$

是分布 $\{q_1,q_2,\cdots,q_r\}$ 的概率生成函数, 进而可以认为 $f_X(x)$ 是复合分布的概率函数.

(b) 证明 S 的矩母函数为

$$M_S(z)=C\{(1-\theta_z)^{-1}\},$$

其中

$$C(z)=\sum_{k=0}^{\infty}c_kz^k=P_N\{Q(z)\}.$$

(c) 如果索赔数分布属于 $(a,b,1)$ 类 (见 4.6.6 节), 描述如何通过递归的方法计算分布 $\{c_k,k=0,1,2,\cdots\}$.

(d) 证明 S 的分布函数为

$$\begin{aligned}F_S(x)&=1-\sum_{n=1}^{\infty}c_n\sum_{j=0}^{n-1}\frac{(x/\theta)^j\mathrm{e}^{-x/\theta}}{j!}\\&=1-\mathrm{e}^{-x/\theta}\sum_{j=0}^{\infty}\bar{C}_j\frac{(x/\theta)^j}{j!},\quad x\geqslant 0,\end{aligned}$$

其中 $\bar{C}_j=\sum\limits_{n=j+1}^{\infty}c_n$.

6.5　计算总索赔额的分布

即使在最简单的情形下, 计算复合模型的分布函数

$$F_S(x)=\sum_{n=0}^{\infty}p_nF_X^{*n}(x),\tag{6.11}$$

或是相应的概率 (密度) 函数都可能是十分复杂的任务. 本节将讨论 (6.11) 式的数值计算方法, 不仅考虑频率和损失分布具体给定的情况, 也考虑其中某个分布或两个分布都是任选的情况.

一种方法是使用近似分布来避免直接计算 (6.11), 这种方法在例 6.5 中已经使用过, 并采用矩方法来估计近似分布的参数. 这种方法的优点是简单, 容易应用, 但其缺点也非常明显. 首先, 没有任何途径可以得知这种近似的优良程度, 选择不同的近似分布可以导致截然不同的结果, 特别是对分布的右尾部. 当然, 近似的优良性可以通过使用更多的矩来增强, 但是四阶矩之后, 分布的信息很快就会用尽.

近似分布也可能无法提供真实分布的某些特征. 例如, 当损失分布是连续型但有一个最大的可能赔付额 (如保单限额) 时, 损失分布也许会在最大值处具有点质量 (也称 "原子"(atom) 或 "峰值"(spike)), 实际的总赔付额分布是混合型分布, 因此, 总赔付额在最大值的整数倍上出现峰值, 相应于出现了 1, 2, 3, $\cdots$ 个最大值的赔案. 如果这些峰值较大, 就可以显著地影响上述最大值的整数倍数点附近的概率. 总赔付额分布的这种跳跃是无法用光滑的近似分布来复制.

第二种计算 (6.11) 式或相应的概率密度函数的方法是直接计算. 其中最困难的 (或应该由计算机完成的) 部分是对 n=2, 3, 4, $\cdots$, 计算损失分布的 n 重卷积. 在某些情形下有解析形式. 例如, 损失程度的分布函数具有卷积封闭性时, 在例 6.15 中有定义, 例 6.12 至例 6.14 给出一些示例. 否则, 卷积必须使用

$$F_X^{*k}(x)=\int_{-\infty}^{\infty} F_X^{*(k-1)}(x-y)\mathrm{d}F_X(y). \tag{6.12}$$

进行数值计算. 当损失被限制在非负数中取值 (通常情形即是如此) 时, 积分限为有限的, (6.12) 式简化为

$$F_X^{*k}(x)=\int_{0-}^{x} F_X^{*(k-1)}(x-y)\mathrm{d}F_X(y). \tag{6.13}$$

这些积分为 Lebesgue-Stieltjes 积分, 因为 $F_X(x)$ 的分布函数可能在 0 点或其他点有跳跃.[①] (6.13) 式的数值计算需要数值积分方法. 积分号中的第一项使得 (6.13) 式要对所有可能的 x 进行计算, 这很快就会使得计算机无法承受.

一种避免这项技术问题的简单方法是在某个便于使用的货币单位 (如 1 000) 的 0, 1, 2, $\cdots$ 倍的点上定义离散分布来取代原有的损失分布, 这将 (6.13) 式化简为 (按新的货币单位)

$$F_X^{*k}(x)=\sum_{y=0}^{x} F_X^{*(k-1)}(x-y)f_X(y).$$

① 这里不探究 Lebesgue-Stieltjes 积分的正式定义, 只需要认识到, 积分 $\int g(y)\mathrm{d}F_X(y)$ 可以用以下方法计算: 对 X 分布的连续点 y 求 $g(y)f_X(y)$ 的积分, 再加上使得 $\Pr(X=y_i)>0$ 的点处 $g(y_i)\Pr(X=y_i)$ 的值. 该形式使得对连续、离散和混合随机变量都可以使用同样的记号.

相应的概率函数为

$$f_X^{*k}(x)=\sum_{y=0}^{x}f_X^{*(k-1)}(x-y)f_X(y).$$

在实际操作中, 货币单位可以制定得足够小以适应最大保险额点尖柱的位置, 我们只需要使最大赔付额是货币单位的整数倍就可以保证尖柱点的位置恰好合适. 随着货币单位的度量变小, 离散分布函数必须越来越接近真实分布函数. 最简单的方法是将每个值都按照一定的货币单位进行取整近似, 例如, 所有的损失或是索赔额都近似到千元. 更为巧妙的方法会在本章后面的内容中讨论.

当损失分布定义在非负整数 0, 1, 2, $\cdots$ 上时, 对整数 x 计算 $f_X^{*k}(x)$ 需要进行 $x+1$ 次乘法. 因此, 要对 $x=0$ 到 $x=n$ 获得公式 (6.11) 的分布, 需要对 n 以内所有可能的 k 值和 x 值进行这样的计算, 该计算过程所需要进行的乘法运算为 n^3 阶, 记作 $\mathrm{O}(n^3)$. 当计算总索赔额分布时用到的索赔额最大值 n 很大时, 计算的次数很快增长, 直至最快的计算机也无法承受. 例如, 真实的索赔案例中 n 值经常高达 1 000, 这就需要大约 10^9 次乘法. 进一步, 若 $\Pr(X=0)>0$, 要获得任何概率都要进行无穷多次计算, 这是因为对所有的 n 和 x, $F_X^{*n}(x)>0$, 所以 (6.11) 式里的求和式包含了无穷多项, 而当 $\Pr(X=0)=0$ 时, 我们有 $F_X^{*n}(x)=0, n>x$. 所以 (6.11) 式最多只有 $x+1$ 个正的项. 表 6-3 给出了后者的一个例子.

下面两节将讨论其他能快速计算总索赔额分布的方法. 第一种方法是**递归法**(the recurseve method), 它可以将上面提到的计算次数减少到 $\mathrm{O}(n^2)$, 这在很大程度上节省了计算机运行时间. 当 n=1 000 时, 同直接计算相比, 递归方法减少了约 99.9% 的计算量. 但是这种方法局限于特定的索赔频率分布, 不过幸运的是, 它包括了所有 4.6 节和附录 B 中讨论的索赔频率分布.

第二种方法是**反演法**(the inversion method), 该方法使用普通的或者专门的反演计算软件包就可以数值地求解逆变换, 如特征函数等. 本章将讨论这种方法的两种具体算法.

6.6 递归方法

假设损失分布 $f_X(x)$ 定义在某个便于使用的货币单位的 $0,1,2,\cdots,m$ 整数倍, m 表示可能的最大赔付额, 可以是无穷大. 进一步假设索赔频率分布 p_k 属于 $(a,b,1)$ 类, 因此满足

$$p_k=\left(a+\frac{b}{k}\right)p_{k-1},\quad k=2,3,4,\cdots.$$

则有下面的结论成立.

定理 6.16 对 $(a,b,1)$ 类, 有

$$f_S(x)=\frac{[p_1-(a+b)p_0]f_X(x)+\sum_{y=1}^{x\wedge m}(a+by/x)f_X(y)f_S(x-y)}{1-af_X(0)},\tag{6.14}$$

注意符号 $x\wedge m$ 表示 $\min(x,m)$.

证明 注意到 $f_X(x)$ 的自变量不能超过 m, 只要适当地替换定理 4.49 中的记号, 这个结果和定理 4.49 是一样的. □

推论 6.17 对 $(a,b,0)$ 类, 结论 (6.14) 可简化为

$$f_S(x)=\frac{\sum_{y=1}^{x\wedge m}(a+by/x)f_X(y)f_S(x-y)}{1-af_X(0)}.\tag{6.15}$$

注意到当损失分布在 0 点没有概率时, (6.14) 式和 (6.15) 式的分母等于 1. 进一步, 在 Poisson 分布情形下, (6.15) 式简化为

$$f_S(x)=\frac{\lambda}{x}\sum_{y=1}^{x\wedge m}yf_X(y)f_S(x-y),\quad x=1,2,\cdots.\tag{6.16}$$

适当地改变定理 4.51 中的记号可得 (6.14) 式和 (6.15) 式递归方法的初始值为 $f_S(0)=P_N[f_X(0)]$. 在 Poisson 分布情形下, 我们有

$$f_S(0)=\mathrm{e}^{-\lambda[1-f_X(0)]}.$$

其他索赔频率分布的初始值见附录 D.

6.6.1 在复合索赔频率模型中的应用

当索赔频率分布可以表示为仅包含 $(a,b,0)$ 类和 $(a,b,1)$ 类分布的复合分布时 (如 Neyman A 型, Poisson 一逆高斯等), 可以两次或多次使用递归公式 (6.14) 来得到总索赔额的分布. 若索赔频率分布可以表示为

$$P_N(z)=P_1[P_2(z)],$$

则总索赔额分布的概率生成函数为

$$P_S(z)=P_N[P_X(z)]=P_1\{P_2[P_X(z)]\},$$

还可以改写为

$$P_S(z)=P_1[P_{S_1}(z)],\tag{6.17}$$

其中

$$P_{S_1}(z)=P_2[P_X(z)].\tag{6.18}$$

现在 (6.18) 式和总损失分布的形式相同. 因此, 若 $P_2(z)$ 属于 $(a,b,0)$ 类或 $(a,b,1)$ 类, 则 S_1 的分布可以用 (6.14) 来计算, 求得的结果就是 (6.18) 式中的 "损失程度" 分布. 因此, 再次将 (6.14) 式应用于公式 (6.17) 就可以得到 S 的分布.

下面的例题将具体说明这个计算过程的使用方法.

例 6.18　赔案数服从 Poisson-ERNB 分布, Poisson 参数为 $\lambda = 2$, ETNB 参数为 $\beta = 3$, $r = 0.2$. 赔付额等于 0, 10, 20 的概率分别为 0.3, 0.5, 0.2. 递归计算总赔付额的分布.

解　沿用前面使用的术语, N 的概率生成函数为 $P_N(z) = P_1[P_2(z)]$, 其中 $P_1(z)$ 和 $P_2(z)$ 分别为 Poisson 和 ETNB 的概率生成函数, 总赔付额分布的概率生成函数为 $P_S(z) = P_1[P_{S_1}(z)]$, 其中 $P_{S_1}(z) = P_2[P_x(z)]$ 为复合 ETNB 分布的概率生成函数. 首先, 我们计算 S_1 的分布. 以 10 为货币单位, 则 $f_X(0) = 0.3$, $f_X(1) = 0.5$, $f_X(2) = 0.2$. 要使用复合 ETNB 的递归, 由

$$\begin{aligned}
f_{S_1}(0) &= P_2[f_X(0)] \\
&= \frac{\{1+\beta[1-f_X(0)]\}^{-r} - (1+\beta)^{-r}}{1-(1+\beta)^{-r}} \\
&= \frac{\{1+3(1-0.3)\}^{-0.2} - (1+3)^{-0.2}}{1-(1+3)^{-0.2}} \\
&= 0.163\ 69.
\end{aligned}$$

开始, 在 (6.14) 式中用 S_1 替代 S, 就可以得到 $f_{S_1}(x)$ 的其他值. 此时有 $a = 3/(1+3) = 0.75$, $b = (0.2-1)a = -0.6$, $p_0 = 0$ 和 $p_1 = (0.2)(3)/[(1+3)^{0.2+1} - (1+3)] = 0.469\ 47$. 则 (6.14) 变为

$$\begin{aligned}
f_{S_1}(x) &= \frac{\begin{array}{l}[0.469\ 47 - (0.75-0.6)(0)]f_X(x) \\ \qquad + \sum_{y=1}^{x}(0.75-0.6y/x)f_X(y)f_{S_1}(x-y)\end{array}}{1-(0.75)(0.3)} \\
&= 0.605\ 77 f_X(x) + 1.290\ 32 \sum_{y=1}^{x}\left(0.75 - 0.6\frac{y}{x}\right) f_X(y) f_{S_1}(x-y).
\end{aligned}$$

最初的几个概率值为

$$\begin{aligned}
f_{S_1}(1) &= 0.605\ 77(0.5) + 1.290\ 32\left[0.75 - 0.6\left(\frac{1}{1}\right)\right](0.5)(0.163\ 69) \\
&= 0.318\ 73,
\end{aligned}$$

$$f_{S_1}(2) = 0.605\ 77(0.2) + 1.290\ 32\left\{\left[0.75 - 0.6\left(\frac{1}{2}\right)\right](0.5)(0.318\ 73)\right.$$
$$\left.+\left[0.75 - 0.6\left(\frac{2}{2}\right)\right](0.2)(0.163\ 69)\right\} = 0.220\ 02,$$
$$f_{S_1}(3) = 1.290\ 32\left\{\left[0.75 - 0.6\left(\frac{1}{3}\right)\right](0.5)(0.220\ 02)\right.$$
$$\left.+\left[0.75 - 0.6\left(\frac{2}{3}\right)\right](0.2)(0.318\ 73)\right\} = 0.106\ 86,$$
$$f_{S_1}(4) = 1.290\ 32\left\{\left[0.75 - 0.6\left(\frac{1}{4}\right)\right](0.5)(0.106\ 86)\right.$$
$$\left.+\left[0.75 - 0.6\left(\frac{2}{4}\right)\right](0.2)(0.220\ 02)\right\} = 0.066\ 92.$$

现在利用复合 Poisson 的概率生成函数来计算 S 的分布,

$$P_S(z) = P_1[P_S(z)] = \mathrm{e}^{\lambda[P_{S_1}(z)-1]}.$$

此时, 在应用复合 Poisson 递归公式时, 分布

$$\{f_{S_1}(x); x = 0, 1, 2, \cdots\}$$

成了 "次级分布" 或 "索赔额分布". 因此, 有

$$f_S(0) = P_S(0) = \mathrm{e}^{\lambda[P_{S_1}(0)-1]} = \mathrm{e}^{\lambda[f_{S_1}(0)-1]} = \mathrm{e}^{2(0.163\ 69-1)} = 0.187\ 75.$$

剩下的概率可以由以下递归公式得到

$$f_S(x) = \frac{2}{x}\sum_{y=1}^{x} y f_{S_1}(y) f_S(x-y), \quad x = 1, 2, \cdots.$$

最初的几个概率值为

$$f_S(1) = 2\left(\frac{1}{1}\right)(0.318\ 73)(0.187\ 75) = 0.119\ 68,$$
$$f_S(2) = 2\left(\frac{1}{2}\right)(0.318\ 73)(0.119\ 68) + 2\left(\frac{2}{2}\right)(0.220\ 02)(0.187\ 75) = 0.120\ 76,$$
$$f_S(3) = 2\left(\frac{1}{3}\right)(0.318\ 73)(0.120\ 76) + 2\left(\frac{2}{3}\right)(0.220\ 02)(0.119\ 68)$$

$$+2\left(\frac{3}{3}\right)(0.106\ 86)(0.187\ 75)=0.100\ 90,$$

$$f_S(4)=2\left(\frac{1}{4}\right)(0.318\ 73)(0.100\ 90)+2\left(\frac{2}{4}\right)(0.220\ 02)(0.120\ 76)$$

$$+2\left(\frac{3}{4}\right)(0.106\ 86)(0.119\ 68)+2\left(\frac{4}{4}\right)(0.066\ 92)(0.187\ 75)$$

$$=0.086\ 96.$$

□

如果重复地应用上述推导, 就可以将这个简单的想法推广到更高阶的复合问题上. 两次应用 (6.14) 式所要求的计算机运行时间大约是一次应用时间的两倍. 但是, 总的计算次数还是 $O(x^2)$ 阶而不是直接计算时的 $O(x^3)$ 阶.

当损失分布的最大可能值为 m 时, 计算速度可以更快, 因为 (6.14) 式中的和式至多有 m 个非零项. 此时, 计算次数可以认为是 $O(x)$ 阶的.

6.6.2 溢出问题

递归式 (6.14) 的初始值为 $P(S=0)=P_N[f_X(0)]$. 对较大的保单组合, 这个概率值非常小, 有时甚至比计算机能够显示的最小值还要小. 如果这种情况发生, 计算机将显示初值为 0, 递归式 (6.14) 无法进行下去. 有几种方法可以克服这个问题 (见 Panjer and Willmot 的文献 [105]), 一种最简单的方法是为 $f_S(0), f_S(1), \cdots, f_S(k)$ 任意设定一组值, 如 $(0, 0, 0, \cdots, 0, 1)$, 其中 k 足够靠左以保证 $F_S(k)$ 仍然是可以忽略的. 可以考虑将 k 值设在分布的均值左边 6 倍标准差的位置就足够了. 利用这个初值, 递归式 (6.14) 可以生成一系列分布的值, 直到一组值一致的小于 $f_S(k)$ 为止. 然后将这些 "概率" 加起来, 并将每个概率值都除以总概率, 从而使得 "真实" 概率的和为 1. 对具体的问题, 可以用试错的方法来确定 k 应取多小.

另一种在初值太小时计算概率的方法是对保单组合按子集进行计算. 例如, 对均值为 λ 的 Poisson 分布, 找一个值 $\lambda^*=\lambda/2^n$, 使得当使用 λ^* 作为 Poisson 分布的均值时 0 点的概率可以在计算机上显示. 这时使用 λ^* 作为 Poisson 分布的均值, 公式 (6.14) 可以用来得到总索赔额的分布. 若 $P_*(z)$ 是使用 λ^* 作为 Poisson 分布的均值时总索赔额的概率生成函数, 那么 $P_S(z)=[P_*(z)]^{2^n}$. 这样一来, 我们就可以通过反复用分布函数与自身作卷积, 逐步得到母函数为 $[P_*(z)]^2, [P_*(z)]^4, [P_*(z)]^8, \cdots, [P_*(z)]^{2^n}$ 的分布. 这需要在计算时额外地进行 n 次卷积, 并且未作任何近似. 这样的过程可以用来计算任何对卷积封闭的分布 —— 对负二项分布, 该过程由 $r^*=r/2^n$ 开始; 对二项分布, 由于 m 必须是整数, 可以进行小幅的修正, 令 $m^*=\lfloor m/2^n\rfloor$, 这里 $\lfloor\bullet\rfloor$ 表示函数的整数部分. 在进行 n 次卷积后, 还需要对参数 $m-m^*2^n$ 使用 (6.14) 式进行计算, 并将这项计算的结果与 n 次卷积的结果再进行卷积. 对复合索赔频率分布, 只需要主分布对卷积封闭.

6.6.3 数值稳定性

任何递归公式都对数值的精度有要求. 因为每个中间计算结果都会用来计算后面的值, 递归方法本身存在的误差在后续的计算中会产生增殖风险, 当计算步骤越来越多时, 误差有 "爆炸" 的可能. 在递归公式 (6.14) 中, 由于在每一步中计算机只能表示有限位的有效数字, 误差就在这样的四舍五入或者说截断中产生了. 稳定性的问题在于: "当存在误差的值被应用在连续的计算过程中时, 误差以多快的速度增长?"

回归公式中的误差增殖问题已经是数值分析学的一个课题, 在 Panjer and Wang[104] 中将这个课题的相关结果推广到对递归公式 (6.14) 的研究上. 这种分析过程非常复杂远远超过了本书的讨论范围, 不过, 可以在这里给出一些主要的结论.

递归公式 (6.14) 中, 误差是在求和

$$\sum_{y=1}^{x}\left(a+\frac{by}{x}\right)f_X(y)f_S(x-y)$$

计算下一个值时产生. 在 S 分布的足够靠右的尾部, 这个数值是正的 (至少非负), 和式的后续值会递减. 若和式中每一项的三个因子都是正的, 即使存在四舍五入误差, 求和的结果仍然是正的. 这种情形下, 递归公式是稳定的, 产生的相对误差不会快速增长. 对基于 Poisson 和负二项的分布, 每一项中的因子始终为正.

另一方面, 对二项分布, 因为 a 为负 b 为正, 且 y/x 是不超过 1 的正数, 所以和式可能会出现负项. 这种情况下, 负项可能会导致正负号在连续的值中交替出现的 "爆炸" 现象, 这种结果显然是荒谬的. 虽然在现实操作中这种情况不常发生, 但读者应该意识到当模型基于二项分布时, 发生这种情况的可能性是存在的.

6.6.4 连续的损失分布

我们已经得到离散损失分布的递归公式, 但人们更惯于使用连续型损失分布. 对连续的损失分布, 递推公式 (6.14) 式将变为一个积分方程, 它的解就是总损失分布.

定理 6.19 *对 $(a,b,1)$ 类索赔频率分布和任何位于正实轴的连续损失分布, 下面的积分方程成立*

$$f_S(x)=p_1f_X(x)+\int_0^x\left(a+\frac{by}{x}\right)f_X(y)f_S(x-y)\mathrm{d}y. \tag{6.19}$$

这个结论的证明超出了本书的范围. 详细的证明过程和相关的推论见 Panjer and Willmot[106] 中的定理 6.14.1 和定理 6.16.1. 他们考虑的是更一般的 (a,b,m)

类分布, 这类分布允许前面的 m 个初值任意选取. 注意, 这时的初始项是 $p_1 f_X(x)$, 而不是 (6.14) 式中的 $[p_1-(a+b)p_0]f_X(x)$. 同样, (6.19) 对 $(a,b,0)$ 类分布也成立.

形如 (6.19) 式的积分方程是第二类 Volterra 积分方程, Baker [10] 讨论了这类积分方程的数值解. 下面我们将对损失分布进行离散化近似然后使用公式 (6.14), 从而避开 Baker [10] 中的复杂计算.

6.6.5 构造算数分布

为了使用递归方法, 最简单的办法是按照某种方便的度量单位 h 进行取整处理, 构造一个离散的损失分布, 称这里的 h 为**跨度**(span). 称这种分布为算术 (arithmetic) 的, 因为它定义在非负整数上. 要算术化一个分布, 重要的是在整个分布的范围内保留原分布的局部性质, 同时还要保留原分布的整体性质, 即保留分布的大致形状, 并且使诸如矩之类的量保持不变.

下面提供的方法可用于离散化 (算术化) 连续分布, 混合分布以及非算术化的离散分布.

舍入 (质量分散) 法

记 jh 处的概率为 f_j, $j=0,1,2,\cdots$, 再令①

$$
\begin{aligned}
f_0 &= \Pr\left(X<\frac{h}{2}\right)=F_X\left(\frac{h}{2}-0\right),\\
f_j &= \Pr\left(jh-\frac{h}{2}\leqslant X<jh+\frac{h}{2}\right)\\
&= F_X\left(jh+\frac{h}{2}-0\right)-F_X\left(jh-\frac{h}{2}-0\right),\quad j=1,2,\cdots.
\end{aligned}
$$

这种方法将 $(j+1)h$ 和 jh 之间的概率分成两个部分并分配到 $j+1$ 和 j 两个点上. 这实际上是将所有的量都舍入到最近的货币单位 h(分布的跨度) 上.

局部矩匹配法

这里要构造的算术分布的前 p 阶矩和原真实的损失分布相匹配. 考虑任意一个长度为 ph 的区间, 记为 $[x_k, x_k+ph)$, 设点 $x_k, x_k+h,\cdots,x_k+ph$ 处的质量为 $m_0^k, m_1^k,\cdots,m_p^k$, 且使得前 p 阶矩不变. 这些条件将构成 $p+1$ 个方程

$$
\sum_{j=0}^{p}(x_k+jh)^r m_j^k=\int_{x_k-0}^{x_k+ph-0} x^r \mathrm{d}F_X(x),\quad r=0,1,2,\cdots,p, \tag{6.20}
$$

其中, 积分上下限中的符号 "−0" 表示积分区间包括 x_k 处的离散概率但不包括 x_k+ph 处的离散概率.

① 记号 $F_X(x-0)$ 表示 x 点的离散概率不包括在内, 对连续分布则没有这样的区别.

适当地排列区间可使得 $x_{k+1}=x_k+ph$, 这样相邻区间的端点重合, 端点的点质量可以相加. 令 $x_0=0$, 得到的离散分布的概率依次为

$$\begin{aligned} &f_0=m_0^0, \quad && f_1=m_1^0, \quad && f_2=m_2^0,\cdots, \\ &f_p=m_p^0+m_0^1, \quad && f_{p+1}=m_1^1, \quad && f_{p+2}=m_2^1,\cdots. \end{aligned} \tag{6.21}$$

令 $x_0=0$, 将 (6.20) 式对所有可能的 k 值求和, 显然整个分布的前 p 阶矩保持不变, 且所有概率的和正好为 1. 下面只需要解方程组 (6.20).

定理 6.20 *方程 (6.20) 的解为*

$$m_j^k=\int_{x_k-0}^{x_k+ph-0}\prod_{i\neq j}\frac{x-x_k-ih}{(j-i)h}\mathrm{d}F_X(x),\quad j=0,1,\cdots,p. \tag{6.22}$$

证明 多项式 $f(y)$ 在点 $y_0,y_1,\cdots,y_n$ 处的 Lagrange 公式为

$$f(y)=\sum_{j=0}^{n}f(y_j)\prod_{i\neq j}\frac{y-y_i}{y_j-y_i}.$$

将这个公式对多项式 $f(y)=y^r$ 应用到点 $x_k,x_k+h,\cdots,x_k+ph$, 得到

$$x^r=\sum_{j=0}^{p}(x_k+jh)^r\prod_{i\neq j}\frac{x-x_k-ih}{(j-i)h},\quad r=0,1,\cdots,p.$$

对 x^r 在区间 $[x_k,x_k+ph)$ 用损失程度的分布函数进行积分, 得到

$$\int_{x_k-0}^{x_k+ph-0}x^r\mathrm{d}F_X(x)=\sum_{j=0}^{p}(x_k+jh)^r m_j^k,$$

其中 m_j^k 由 (6.22) 式给出. 这样, (6.22) 式的解就保持了前 p 阶矩不变, 符合要求.□

例 6.21 *设 X 服从指数分布, 概率密度函数为 $f(x)=0.1\mathrm{e}^{-0.1x}$. 用舍入法和一阶矩匹配法来离散化此分布, 跨度 h=2.*

解 舍入法的一般公式为

$$\begin{aligned} f_0&=F(1)=1-\mathrm{e}^{-0.1(1)}=0.095\,16, \\ f_j&=F(2j+1)-F(2j-1)=\mathrm{e}^{-0.1(2j-1)}-\mathrm{e}^{-0.1(2j+1)}. \end{aligned}$$

表 6-9 给出了前几个取值.

表 6-9 用两种方法离散化的指数分布

j	舍入法的 f_j	匹配法的 f_j
0	0.095 16	0.093 65
1	0.164 02	0.164 29
2	0.134 29	0.134 51
3	0.109 95	0.110 13
4	0.090 02	0.090 17
5	0.073 70	0.073 82
6	0.060 34	0.060 44
7	0.049 40	0.049 48
8	0.040 45	0.040 51
9	0.033 11	0.033 17
10	0.027 11	0.027 16

使用一阶矩匹配法, 我们有 $p=1$ 和 $x_k=2k$, 关键方程为

$$m_0^k=\int_{2k}^{2k+2}\frac{x-2k-2}{-2}(0.1)\mathrm{e}^{-0.1x}\mathrm{d}x=5\mathrm{e}^{-0.1(2k+2)}-4\mathrm{e}^{-0.1(2k)},$$

$$m_1^k=\int_{2k}^{2k+2}\frac{x-2k}{2}(0.1)\mathrm{e}^{-0.1x}\mathrm{d}x=-6\mathrm{e}^{-0.1(2k+2)}+5\mathrm{e}^{-0.1(2k)},$$

从而

$$f_0=m_0^0=5\mathrm{e}^{-0.2}-4=0.093\ 65,$$

$$f_j=m_1^{j-1}+m_0^j=5\mathrm{e}^{-0.1(2j-2)}-10\mathrm{e}^{-0.1(2j)}+5\mathrm{e}^{-0.1(2j+2)}.$$

表 6-9 也给出了此方法的前几个值. 对一阶矩匹配法, 习题 6.36 给出了一个更直接的解法. □

这种局部矩匹配方法最先由 Gerber and Jones[44] 以及 Gerber[45] 提出, Panjer and Lutek[103] 对多种经验的和解析的损失分布作了进一步的研究. 在评估误差对总止损净保费 (总超额损失净保费) 的影响时, Panjer and Lutek[103] 指出两阶矩通常就足够了, 增加第三阶矩的匹配条件只能微小地提高精确度. 此外, 舍入法和一阶矩匹配法 ($p=1$) 具有非常接近的误差, 而二阶矩匹配法 ($p=2$) 有明显地改进. 附录 E 给出了舍入法和一阶矩匹配法的详细公式. 人们愿意选择全概率的匹配或仅匹配一阶矩的重要原因为, 这样得到的概率总是非负的, 当需要匹配两阶或更高阶矩时, 不能保证这一点.

这里描述的方法本质上类似于数值分析中用来解决形如 (6.19) 式的 Volterra 积分方程的数值方法, 见 Baker[10].

习题

6.36 证明局部矩匹配法, 当 $k=1$(即匹配总概率和均值) 时, 由 (6.21) 式和 (6.22) 式可得

$$f_0 = 1 - \frac{\mathrm{E}[X \wedge h]}{h},$$
$$f_i = \frac{2\mathrm{E}[X \wedge ih] - \mathrm{E}[X \wedge (i-1)h] - \mathrm{E}[X \wedge (i+1)h]}{h}, \quad i = 1, 2, \cdots,$$

且 $\{f_i; i = 0, 1, 2, \cdots\}$ 的确为一个有效的分布, 其均值与原损失分布相同. 使用这里给出的公式验证例 6.21 中的公式.

6.37 你是某棒球选手的经纪人, 需要为该选手签订一份奖金合同, 具体如表 6-10 所示. 已知, 轮击的次数服从参数 $k=200$ 的 Poisson 分布, 参数 x 应满足: 该选手至少挣 4 000 000 美元的概率等于 0.95. 求这位选手的奖金期望值.

表 6-10 习题 6.37 的数据

表击打种类	击中的概率	每一击的赔付额
单击	0.14	x
双杀	0.05	$2x$
三杀	0.02	$3x$
本垒打	0.03	$4x$

6.38 有些作者用下面的两个 Poisson 的加权平均来区分 "好" 司机和 "坏" 司机 (见例 4.64), 如 Tröbliger[130] 中的做法,

$$p_k = w\frac{\mathrm{e}^{-\lambda_1}\lambda_1^k}{k!} + (1-w)\frac{\mathrm{e}^{-\lambda_2}\lambda_2^k}{k!}.$$

(a) 用两种类型司机的损失次数的概率生成函数 $P_1(z)$ 和 $P_2(z)$ 来表示总损失次数的概率生成函数 $P_N(z)$.

(b) 令 $f_X(x)$ 为定义在非负整数上的损失分布, 如何使用 (6.16) 式来计算整个团体的总赔付额分布?

(c) 是否可以将这种方法推广到其他索赔频率分布?

6.39 某复合 Poisson 总损失模型每年的期望赔案数为 5, 损失分布以 1 000 为单位, 已知 $f_S(1) = \mathrm{e}^{-5}$ 和 $f_S(2) = \frac{5}{2}\mathrm{e}^{-5}$, 求 $f_X(2)$.

6.40 已知某复合 Poisson 分布的参数 $\lambda = 6$ 且个体损失额的概率函数为 $f_X(1) = f_X(2) = f_X(4) = \frac{1}{3}$. 表 6-11 给出了总损失分布 S 的一些概率函数值, 求 $f_S(6)$.

6.41 考虑 $(a,b,0)$ 类损失频率分布和任何定义在正整数 $\{1, 2, \cdots, M < \infty\}$ 上的损失分布, 其中 M 是单次损失的最大值.

(a) 证明对复合分布, 有下面的回溯公式成立：

$$f_S(x)=\frac{f_S(x+M)-\sum_{y=1}^{M-1}\left(a+b\frac{M-y}{x+M}\right)f_X(M-y)f_S(x+y)}{\left(a+b\frac{M}{x+M}\right)f_X(M)}.$$

(b) 对参数为 (m,q) 的二项索赔频率分布, 怎样利用上述公式得到总损失的分布? 见 Panjer and Wang[104].

表 6-11 习题 6.40 的数据

x	$f_S(x)$
3	0.013 2
4	0.021 5
5	0.027 1
6	$f_S(6)$
7	0.041 0

6.42 总赔付额服从参数 $\lambda=2$ 的复合 Poisson 分布, $f_X(1)=1/4$, $f_X(2)=3/4$. 保费为 6, 保险人不仅承保总索赔额, 并同意支付红利 (保费返还), 红利等于保费的 75%减去索赔额的 100%(若还有剩余). 求保费扣除赔付额和分红的期望后的余额.

6.43 某医生一天中为 N_A 位成人和 N_C 位儿童提供医疗服务, 假设 N_A 和 N_C 分别服从参数为 3 和 2 的 Poisson 分布. 每位病人的治疗时间分布如下表所示：

	成人	儿童
1 小时	0.4	0.9
2 小时	0.6	0.1

令 N_A, N_C 以及所有个体的治疗时间长短是相互独立的, 医生对病人的医疗服务每小时收费 200 元, 求诊所一天内收入小于或等于 800 的概率.

6.44 某团体保单的总索赔额 S 服从参数 $\lambda=1$ 的复合 Poisson 分布, 所有的索赔额都等于 2, 保险人支付的红利如下：

$$D=\begin{cases}6-S, & S<6,\\ 0, & S\geqslant 6.\end{cases}$$

求 $\mathrm{E}[D]$.

6.45 给定两个独立的复合 Poisson 随机变量 S_1 和 S_2, 其中 $f_j(x), j=1,2$ 是两个单次赔案索赔额的分布, $\lambda_1=\lambda_2=1$, $f_1(1)=1$, $f_2(1)=f_2(2)=0.5$. 令 $F_X(x)$ 是复合分布 $S=S_1+S_2$ 的单次赔案的索赔额的分布函数, 计算 $F_X^{*4}(6)$.

6.46 变量 S 为复合 Poisson 索赔分布, 满足以下性质：

(1) 个体索赔额等于 1, 2 或 3; (2) $\mathrm{E}(S)=56$;

(3) $\mathrm{Var}(S) = 126$; **(4)** $\lambda = 29$.

求索赔额等于 2 的赔案个数的期望值.

6.47 对索赔额为正整数的复合 Poisson 分布, 概率函数满足

$$f_S(x) = \frac{1}{x}[0.16f_S(x-1) + kf_S(x-2) + 0.72f_S(x-3)], \quad x = 1, 2, 3, \cdots.$$

总赔付额的期望值为 1.68, 求赔案数的期望值.

6.48 已知某保单组合如下:

(1) 赔案数服从 Poisson 分布;

(2) 索赔额为 1, 2 或 3;

(3) 对不同免赔额的止损再保险合同的净保费如表 6-12 所示.

求总损失额为 5 或 6 的概率.

表 6-12 习题 6.48 的数据

免赔额	净保费
4	0.20
5	0.10
6	0.04
7	0.02

6.49 对团体失能收入保险, 每年每 100 个投保人发生伤残的期望人数为 1 人. 每次伤残的持续天数 Y 的 (生存) 函数为

$$\Pr(Y > y) = 1 - \frac{y}{10}, \quad y = 0, 1, \cdots, 10.$$

在五天等待期后, 赔付额为每日 20 元. 使用复合 Poisson 分布, 求由 1 500 人构成的团体的总损失额的方差.

6.50 某总体中有两类司机, 已知每位司机的年事故数服从几何分布. 从类型 I 中随机挑选的司机, 几何分布的参数服从区间 (0, 1) 上的均匀分布, 25% 的司机属于类型 I, 所有类型 II 中的司机的期望赔案数都是 0.25. 对于从总体中随机挑选的司机, 求每年恰好发生两次事故的概率.

提示: 下面两题需要计算机编程计算.

6.51 某保单承保了某公司车队的卡车造成的所有财产损失. 一年内的损失次数服从参数 $\lambda = 5$ 的 Poisson 分布, 单次损失额服从参数 $\alpha = 0.5$, $\theta = 2\,500$ 的 gamma 分布, 该合同每年的最大赔付额为 20 000. 使用跨度为 100 的舍入法, 求赔付额等于最大赔付额的概率.

6.52 某人购买了一份健康保险. 已知每次医生出诊须支付 10 元, 每个药方须支付 5 元. 支付 10 元的概率为 0.25, 支付 5 元的概率为 0.75. 总赔付次数服从参数 $\lambda_1 = 10, \lambda_2 = 4$ 的 Poisson-Poisson(Neyman A 型) 分布. 求一年内总赔付额超过 400 的概率, 并与正态分布近似的结果进行比较.

6.53 证明若用舍入法离散化指数分布, 得到的离散分布为 ZM 几何分布.

6.7 个体保单的更改对总赔付额的影响

在 5.6 节中, 讨论了个体免赔额 (普通的或特权的) 同时影响个体索赔额和索赔频率分布的现象. 本节将考虑免赔额对总损失的影响. 值得注意的是, 个体共保和个体保单限额都会对个体损失额产生影响, 但不影响损失的发生频率, 因此本节将注意力集中在免赔额的问题上. 同时应当注意到, 我们仍然假设对保单的更改并不能对投保人群体的风险特性产生影响, 因此也不可能对保单涉及的个体损失分布产生影响, 这是在 5.6 节中讨论过的问题. 也就是说, 假设个体损失额 X 的 "基础" 分布不受保单更改的影响, 仅仅是赔付额本身受影响.

从总损失额的角度看, 也有一些相关的事实需要注意. 无论免赔额是普通或是特权类型的, 都可以假设个体损失导致赔案发生的概率为 v, 不过要对个体损失的基础随机变量 X 进行相应的调整 (包括免赔额等) 后才决定赔付额. 个体赔付可以看成是基于每次损失为基础, 只不过若损失的发生没有造成赔付的发生, 赔付额就等于 0. 这种赔付额记为 Y^L. 这样, 在每次损失的基础上, 赔付额仍是由每次损失额决定的. 另外, 个体赔付也可以看成是基于每次赔付的, 此时, 赔付额记为 Y^P. 在此基础上, 赔付额仅由那些造成实际的非零赔付发生的损失决定. 因此, 由定义, $\Pr(Y^P=0)=0$, Y^P 的分布为给定 $Y^L>0$ 时 Y^L 的条件分布, 将此关系形式上写成 $Y^P=Y^L|Y^L>0$. 此时, 两者的累积分布函数有如下关系

$$F_{Y^L}(y)=(1-v)+vF_{Y^P}(y),\quad y\geqslant 0,$$

这是因为 $1-v=\Pr(Y^L=0)=F_{Y^L}(0)$(注意到即使 X 以及 Y^P 和 Y^L 在 $y>0$ 处有连续的概率密度函数, Y^L 也在 0 点有离散概率质量 $1-v$). 因此, Y^P 和 Y^L 的矩母函数有如下关系

$$M_{Y^L}(t)=(1-v)+vM_{Y^P}(t),\tag{6.23}$$

由此可得期望值的关系

$$\mathrm{E}(\mathrm{e}^{tY^L})=\mathrm{E}(\mathrm{e}^{tY^L}|Y^L=0)\Pr(Y^L=0)+\mathrm{E}(\mathrm{e}^{tY^L}|Y^L>0)\Pr(Y^L>0).$$

由 5.6 节的讨论可知, 损失次数 N^L 和赔付次数 N^P 的概率生成函数之间的关系为

$$P_{N^P}(z)=P_{N^L}(1-v+vz),\tag{6.24}$$

其中 $P_{N^P}(z)=\mathrm{E}(z^{N^P}), P_{N^L}(z)=\mathrm{E}(z^{N^L})$.

现在回到对总赔付额的分析上. 基于每次损失, 总赔付额可以表示为 $S=Y_1^L+Y_2^L+\cdots+Y_{N^L}^L$, 若 $N^L=0$ 则 $S=0$, 其中 Y_j^L 是对第 j 次损失的赔付额. 另

外, 忽略那些没有造成赔付的损失, 可以基于每次赔付, 将总赔付额表示为 $S = Y_1^P + Y_2^P + \cdots + Y_{N^L}^P$, 若 $\mathrm{N^P} = 0$ 则 $S = 0$, 其中 Y_j^P 是第 j 次造成非零赔付的损失造成的赔付额. 显然, S 可以在加总的原则下有两种不同的表示方式. 首先, 基于每次损失, S 的矩母函数为

$$M_S(t) = \mathrm{E}(\mathrm{e}^{tS}) = P_{N^L}[M_{Y^L}(t)], \tag{6.25}$$

另一方面, 基于每次赔付, 我们有

$$M_S(t) = \mathrm{E}(\mathrm{e}^{tS}) = P_{N^P}[M_{Y^P}(t)]. \tag{6.26}$$

显然, (6.25) 式和 (6.26) 式是相等的, 这可以从 (6.23) 式和 (6.24) 式中看出, 即

$$P_{N^L}[M_{Y^L}(t)] = P_{N^L}[1 - v + vM_{Y^P}(t)] = P_{N^P}[M_{Y^P}(t)].$$

因此, 任何对总赔付额 S 的分析都可以基于每次损失 (矩母函数的复合表述为 (6.25)) 或基于每次赔付 (矩母函数的复合表述为 (6.26)) 来完成. 应该选择哪种分析方法, 显然是由哪种方法更适合手中的具体问题决定的. 若无法看出哪种方法更好, 研究者们发现基于每次损失来估计 S 的矩更加方便. 特别地, 5.5 节中给出的个体均值和方差公式都是基于每次损失的, 总赔付额 S 的均值和方差可以由这些公式结合 (6.6) 式来计算, 只要将 N 替换为 N^Y, X 替换为 Y^L.

另一方面, 如果关心的是 S 的 (近似) 分布, 则一般情况下基于每次赔付来计算更合适, 原因是当 $\mathrm{E}(N^L)$ 很大时, 基于每次损失的计算可能出现下溢出的问题, 特别是使用特权免赔额时, 由于 Y^L 的分布存在大量 0 概率值, 可能会发生计算机存储的问题. 另外, 从方便的角度讲, 也应该先选择依据保单对个体损失进行调整, 再离散化 (若有必要的话), 而非先离散化再将保单的调整方案应用到离散的分布. 尽管如些, 这个问题仅当免赔额和保单限额不是离散跨度的整数倍时才发生. 下面的例子将说明上述的想法.

例 6.22 基础损失发生的次数服从参数 $\lambda = 3$ 的 Poisson 分布, 个体损失服从参数为 $\alpha = 4, \theta = 10$ 的 Pareto 分布, 个体的普通免赔额为 6, 共保比率为 75%, 个体损失限额为 24(在应用免赔额和再保险前). 求总赔付额的均值, 方差和分布.

解 首先基于每次损失计算均值和方差, 损失次数的期望值为 $\mathrm{E}(N^L) = 3$, 基于每次损失的个体赔付均值为 (在定理 5.13 中取 Pareto 分布以及 $r = 0$)

$$\mathrm{E}(Y^L) = 0.75[\mathrm{E}(X \wedge 24) - \mathrm{E}(X \wedge 6)] = 0.75(3.248\ 5 - 2.519\ 5) = 0.546\ 75.$$

因此, 总赔付额的均值为

$$\mathrm{E}(S) = \mathrm{E}(N^L)\mathrm{E}(Y^L) = (3)(0.546\ 75) = 1.64.$$

在定理 5.14 中取 Pareto 分布以及 $r = 0$, 基于每次损失计算的个体赔付额的二阶矩为

$$\begin{aligned}\mathrm{E}[(Y^L)^2] &= (0.75)^2\{\mathrm{E}[(X\wedge 24)^2] - \mathrm{E}[(X\wedge 6)^2] \\ &\quad -2(6)\mathrm{E}(X\wedge 24) + 2(6)\mathrm{E}(X\wedge 6)\} \\ &= (0.75)^2[26.379\,0 - 10.546\,9 - 12(3.248\,5) + 12(2.519\,5)] \\ &= 3.984\,81.\end{aligned}$$

要计算总赔付额的方差, 并不需要确切地求得 $\mathrm{Var}(Y^L)$, 由于 S 为复合 Poisson 分布, 因此 (例如, 使用 (4.27) 式)

$$\mathrm{Var}(S) = \lambda\mathrm{E}[(Y^L)^2] = 3(3.984\,81) = 11.954\,4 = (3.46)^2.$$

要计算 S 的 (近似) 分布, 我们基于每次赔付来考虑. 首先注意到 $v = \Pr(X > 6) = [10/(10+6)]^4 = 0.152\,59$, 赔付次数 N^P 服从 Poisson 分布, 均值为 $\mathrm{E}(N^P) = \lambda v = 3(0.152\,59) = 0.457\,76$. 令 $Z = X - 6|X > 6$, 则 Z 是个体赔付额随机变量, 免赔额为 6. 此时

$$\Pr(Z > z) = \frac{\Pr(X > z+6)}{\Pr(X > 6)}.$$

由于共保比率为 75%, $Y^P = 0.75Z$ 的累积分布函数为

$$F_{Y^P}(y) = 1 - \Pr(0.75Z > y) = 1 - \frac{\Pr(X > 6 + y/0.75)}{\Pr(X > 6)}.$$

即对小于最大赔付额 (0.75)(24−6)=13.5 的 y 值,

$$F_{Y^P}(y) = \frac{\Pr(X > 6) - Pr(X > 6 + y/0.75)}{Pr(X > 6)}, \quad y < 13.5,$$

且对 $y \geqslant 13.5$ 有 $F_{Y^P}(y) = 1$. 接着用跨度 2.25 和舍入法离散化 Y^P 的分布 (注意这里的做法是先应用保单的更改再离散化), 得到 $f_0 = F_{Y^P}(1.125) = 0.301\,24$, $f_1 = F_{Y^P}(3.375) - F_{Y^P}(1.125) = 0.327\,68$ 等, 这种情形下在计算 f_6 时应当注意, 我们有 $f_6 = F_{Y^P}(14.625) - F_{Y^P}(12.375) = 1 - 0.941\,26 = 0.058\,74$, 进而对 $n = 7, 8, \cdots$, $f_n = 1 - 1 = 0$. S 的近似分布可以使用复合 Poisson 的递推公式来计算, 也就是 $g_0 = \mathrm{e}^{-0.457\,76(1-0.301\,24)} = 0.726\,25$,

$$g_k = \frac{0.457\,76}{k}\sum_{j=1}^{k} j f_j g_{k-j}, \quad k = 1, 2, 3, \cdots.$$

因此, 有 $g_1 = (0.457\,76)(1)(0.327\,68)(0.726\,25) = 0.108\,94$. □

习题

6.54 假设基础损失发生的次数 N^L 的概率生成函数为 $P_{N^L}(z)=B[\theta(z-1)]$, 这里 θ 为参数, B 是独立于 θ 的函数. 个体的基础损失分布为指数分布, 累积分布函数为 $F_X(x)=1-e^{-\mu x}, x\geqslant 0$. 个体损失有普通免赔额 d, 分保比例为 α. 证明基于每次赔付, 总赔付额的矩母函数有形如 (6.26) 式的复合形式, 其中 N^P 与 N^L 同分布, 但 θ 要换成 $\theta e^{-\mu d}$, Y^P 与 X 同分布, 但参数 μ 要换成 μ/α.

6.55 个体基础损失的模型服从参数 $\alpha=2, \theta=100$ 的 gamma 分布, 损失次数服从参数 $r=2$, $\beta=1.5$ 的负二项分布. 每个个体损失的普通免赔额为 50, 保单限额为 175(免赔前)

(a) 基于每次损失, 求总赔付额的均值和方差.

(b) 求赔付次数的分布.

(c) 在赔付已发生的条件下, 求每次赔付额 Y^P 的累积分布函数.

(d) 使用舍入法, 令跨度为 40, 离散化 (c) 中的损失分布.

(e) 使用递归公式来计算离散化的总赔付额分布在离散赔付额 120 之前的值.

6.8 近似分布的计算

只要对损失分布采取近似计算, 所得结果必然是对真实总量分布的近似. 特别地, 真实的总分布通常是连续的 (除了可能在 0 点或在总删失限额点有离散概率外), 而近似分布要么就像在使用递归式或者快速傅里叶变换 (FFT) 时那样, 在离散的等间距的点上有概率; 要么就像在进行随机模拟时那样, 在任意值上有离散概率 $1/n$; 要么就像在使用 Heckman-Meyers 方法时那样, 是分段线性的分布函数. 本节将介绍几种通过这些近似分布来计算 $F_S(x)$ 和 $E[(S\wedge x)^k]$ 的合理方法. 始终假设总赔付额的真实分布是连续的, 只可能在 S=0 处有离散概率.

6.8.1 算术分布

对递归法和 FFT 法, 近似分布可以写成 $p_0, p_1, \cdots$, 其中 $p_j=\Pr(S^*=jh)$, S^* 表示近似分布. 有几种可行的连续化分布的方法, 我们介绍其中的一种. 假设能得到总赔付额为 0 的真实概率 $g_0=\Pr(S=0)$. 该方法的基本想法是, 对 $j=1,2,\cdots$, 假设概率 p_j 是均匀遍布在 $(j-1/2)h$ 到 $(j+1/2)h$ 的区间上, 进而构造 S^* 的连续近似. 对从 0 到 $h/2$ 的区间, 离散概率 g_0 被放置在 0 点, 剩下的概率 p_0-g_0, 均匀地分布在该区间上. 令 S^{**} 是服从这种混合分布的随机变量, 所有我们关心的量都采用 S^{**} 计算.

例 6.23 *令 N 服从参数 $\beta=2$ 的几何分布, X 服从参数 $\theta=100$ 的指数分布. 使用跨度为 2 的递归法来近似总分布, 并给出一个连续近似.*

解 指数分布可以通过匹配一阶矩的方法来离散化, 表 6-13 中给出概率. 表 6-13 还给出了使用递归法计算得到的总概率. 注意到 $g_0=\Pr(N=0)=(1+\beta)^{-1}=1/3$,

对 $j=1,2,\cdots$, 连续近似分布的概率密度函数为 $f_{S^{**}}(x)=f_{S^*}(2j)/2$, $2j-1<x\leqslant 2j+1$. 进而 $\Pr(S^{**}=0)=1/3$, $f_{S^{**}}(x)=(0.335\ 556-1/3)/1=0.002\ 223$, $0<x\leqslant 1$.

表 6-13　总赔付额分布的离散近似

j	x	$f_X(x)$	$p_j=f_{S^*}(x)$
0	0	0.009 934	0.335 556
1	2	0.019 605	0.004 415
2	4	0.019 216	0.004 386
3	6	0.018 836	0.004 356
4	8	0.018 463	0.004 327
5	10	0.018 097	0.004 299
6	12	0.017 739	0.004 270
7	14	0.017 388	0.004 242
8	16	0.017 043	0.004 214
9	18	0.016 706	0.004 186
10	20	0.016 375	0.004 158

回到初始的问题, 现在可以给出这些基本量的一般公式. 对累积分布函数

$$\begin{aligned} F_{S^{**}}(x) &= g_0+\int_0^x \frac{p_0-g_0}{h/2}\mathrm{d}s \\ &= g_0+\frac{2x}{h}(p_0-g_0),\quad 0\leqslant x\leqslant \frac{h}{2}, \end{aligned}$$

$$\begin{aligned} F_{S^{**}}(x) &= \sum_{i=0}^{j-1} p_i+\int_{(j-1/2)h}^x \frac{p_j}{h}\mathrm{d}s \\ &= \sum_{i=0}^{j-1} p_i+\frac{x-(j-1/2)h}{h}p_j,\quad \left(j-\frac{1}{2}\right)h<x\leqslant\left(j+\frac{1}{2}\right)h. \end{aligned}$$

带限额的期望值 (LEV) 为

$$\begin{aligned} \mathrm{E}[(S^{**}\wedge x)^k] &= 0^k g_0+\int_0^x s^k\frac{p_0-g_0}{h/2}\mathrm{d}s+x^k[1-F_{S^{**}}(x)] \\ &= \frac{2x^{k+1}(p_0-g_0)}{h(k+1)}+x^k[1-F_{S^{**}}(x)],\quad 0<x\leqslant\frac{h}{2}, \end{aligned}$$

$$\begin{aligned} \mathrm{E}[(S^{**}\wedge x)^k] = {}& 0^k g_0+\int_0^{h/2} s^k\frac{p_0-g_0}{h/2}\mathrm{d}s+\sum_{i=1}^{j-1}\int_{(i-1/2)h}^{(i+1/2)h} s^k\frac{p_i}{h}\mathrm{d}s \\ &+\int_{(j-1/2)h}^x s^k\frac{p_j}{h}\mathrm{d}s+x^k[1-F_{S^{**}}(x)] \end{aligned}$$

$$= \frac{(h/2)^k(p_0 - g_0)}{k+1} + \sum_{i=1}^{j-1} \frac{h^k[(i+1/2)^{k+1} - (i-1/2)^{k+1}]}{k+1} p_i$$
$$+ \frac{x^{k+1} - [(j-1/2)h]^{k+1}}{h(k+1)} p_j$$
$$+ x^k[1 - F_{S^{**}}(x)], \qquad \left(j - \frac{1}{2}\right)h < x \leqslant \left(j + \frac{1}{2}\right)h.$$

$k = 1$ 时, 可以化简为

$$\mathrm{E}(S^{**} \wedge x) = \begin{cases} x(1-g_0) - \dfrac{x^2}{h}(p_0 - g_0), \quad 0 < x \leqslant \dfrac{h}{2}, \\ \dfrac{h}{4}(p_0 - g_0) + \displaystyle\sum_{i=1}^{j-1} ihp_i + \dfrac{x^2 - [(j-1/2)h]^2}{2h} p_j \\ \qquad + x[1 - F_{S^{**}}(x)], \quad \left(j - \dfrac{1}{2}\right)h < x \leqslant \left(j + \dfrac{1}{2}\right)h. \end{cases} \tag{6.27}$$

附录 E 总结了这些公式. □

例 6.24(续例 6.23) 使用 S^*, S^{**} 和总损失的精确分布来计算分布函数和 LEV 在 1 到 10 的整数上的取值.

解 此例可以得到精确分布. 由例 6.12 可知 $\Pr(S = 0) = (1+\beta)^{-1} = 1/3$, 连续部分的概率密度函数为

$$f_S(x) = \frac{\beta}{\theta(1+\beta)^2} \exp\left[-\frac{x}{\theta(1+\beta)}\right] = \frac{2}{900}\mathrm{e}^{-x/300}, \quad x > 0.$$

由此有

$$F_S(x) = \frac{1}{3} + \int_0^x \frac{2}{900}\mathrm{e}^{-s/300}\mathrm{d}s = 1 - \frac{2}{3}\mathrm{e}^{-x/300},$$

$$\mathrm{E}(S \wedge x) = \int_0^x \frac{2s}{900}\mathrm{e}^{-s/300}\mathrm{d}s + x\frac{2}{3}\mathrm{e}^{-x/300} = 200(1 - \mathrm{e}^{-x/300}).$$

待求的值在表 6-14 中给出. □

6.8.2 经验分布

如果近似分布是由随机模拟得到的 (模拟过程将在第 17 章中讨论), 其结果就是经验分布. 不同于由递归法或是 FFT 得出的近似, 随机模拟不会将概率分配在等间距的点上, 这导致平滑近似分布的方法并不是一眼就能看出. 另一方面, 模拟通常会包括数万或数十万个点, 因此各个点互相靠得很近. 基于以上原因, 可以简单地使用经验分布作为问题的答案, 也就是说, 所有的计算都通过近似经验随机变

量 S^* 来完成. 一般情况下的计算公式都非常简单. 令 $x_1, x_2, \cdots, x_n$ 为模拟值, 那么

$$F_{S^*}(x) = \frac{x_j \leqslant x \text{ 的数目}}{n},$$

$$\mathrm{E}[(S^* \wedge x)^k] = \frac{1}{n} \sum_{x_j < x} x_j^k + x^k[1 - F_{S^*}(x)].$$

表 6-14 真实总赔付额的分布函数和两种近似的分布函数的比较

x	cdf			LEV		
	S	S^*	S^{**}	S	S^*	S^{**}
1	0.335 552	0.335 556	0.335 556	0.665 56	0.664 44	0.665 56
2	0.337 763	0.339 971	0.337 763	1.328 90	1.328 89	1.328 90
3	0.339 967	0.339 971	0.339 970	1.990 03	1.988 92	1.990 03
4	0.342 163	0.344 357	0.342 163	2.648 97	2.648 95	2.648 96
5	0.344 352	0.344 357	0.344 356	3.305 71	3.304 59	3.305 70
6	0.346 534	0.348 713	0.346 534	3.960 27	3.960 23	3.960 25
7	0.348 709	0.348 713	0.348 712	4.612 64	4.611 52	4.612 63
8	0.350 876	0.353 040	0.350 876	5.262 85	5.262 81	5.262 84
9	0.353 036	0.353 040	0.353 039	5.910 89	5.909 77	5.910 88
10	0.355 189	0.357 339	0.355 189	6.556 78	6.556 73	6.556 76

例 6.25 (续例 6.23) 从几何损失频率分布和指数损失分布的复合模型模拟 1000 个观测值, 使用所得结果来计算累积分布函数和 LEV 在 1 到 10 的整数上的取值. 这里选择较小的样本容量是为了使得在 0 到 10 之间 (不包括 0) 的模拟个数大约只有 30.

解 模拟过程产生总赔付为 0 的次数为 331, 小于 10 的非零赔付额和第一个大于 10 的赔付额如表 6-15 所示. 除了 0 以外, 模拟过程中没有相同的值出现. 由经验分布得到的值和真实值的比较见表 6-16. □

6.8.3 分段线性累积分布函数

当使用 Heckman-Meyers 反演法 (将在 6.9.2 节中介绍) 时, 得到的是累积分布函数 $F_S(x)$ 在任意给定集合上的近似值, 其误差来自于对损失分布函数的分段线性要求以及近似积分的使用. 设 $0 = x_1 < x_2 < \cdots < x_n$ 是任意选定的点, $S^\#$ 为任意选定的随机变量, 且满足在点 $x_1, x_2, \cdots, x_n$ 上的分布函数值与 Heckman-Meyers 方法的结果相一致. 令 $F_j = F_{S^\#}(x_j)$, 再令 $F_n = 1$ 以保证没有概率丢失. 构造光滑化分布的最简单的方法是用直线将这些点联结起来. 记 $S^{\#\#}$ 是具有这种特殊累积分布函数的随机变量, $S^{\#\#}$ 的累积分布函数的中间值由插值得到

$$F_{S^{\#\#}}(x) = \frac{(x - x_{j-1})F_j + (x_j - x)F_{j-1}}{x_j - x_{j-1}}.$$

表 6-15 总损失的模拟值

j	x_j	j	x_j
1～331	0	346	6.15
332	0.04	347	6.26
333	0.12	348	6.58
334	0.89	349	6.68
335	1.76	350	6.71
336	2.16	351	6.82
337	3.13	352	7.76
338	3.40	353	8.23
339	4.38	354	8.67
340	4.78	355	8.77
341	4.95	356	8.85
342	5.04	357	9.18
343	5.07	358	9.88
344	5.81	359	10.12
345	5.94		

表 6-16 随机模拟的经验值和光滑分布函数求出的值的比较

x	$F_{S^*}(x)$	$F_S(x)$	$\mathrm{E}(S^* \wedge x)$	$\mathrm{E}(S \wedge x)$
0	0.331	0.333	0.000 0	0.000 0
1	0.334	0.336	0.667 1	0.665 6
2	0.335	0.338	1.332 8	1.328 9
3	0.336	0.340	1.997 0	1.990 0
4	0.338	0.342	2.659 5	2.649 0
5	0.341	0.344	3.320 6	3.305 7
6	0.345	0.347	3.977 5	3.960 3
7	0.351	0.349	4.629 7	4.612 6
8	0.352	0.351	5.278 4	5.262 9
9	0.356	0.353	5.925 0	5.910 9
10	0.358	0.355	6.568 0	6.556 8

带限额的期望值公式 (对 $x_{j-1} < x \leqslant x_j$) 为

$$\begin{aligned}
\mathrm{E}[(S^{\#\#} \wedge x)^k] &= \sum_{i=2}^{j-1} \int_{x_{i-1}}^{x_i} s^k \frac{F_i - F_{i-1}}{x_i - x_{i-1}} \mathrm{d}s \\
&\quad + \int_{x_{j-1}}^{x} s^k \frac{F_j - F_{j-1}}{x_j - x_{j-1}} \mathrm{d}s + x^k[1 - F_{S^{\#\#}}(x)] \\
&= \sum_{i=2}^{j-1} \frac{(x_i^{k+1} - x_{i-1}^{k+1})(F_i - F_{i-1})}{(k+1)(x_i - x_{i-1})} \\
&\quad + \frac{(x^{k+1} - x_{j-1}^{k+1})(F_j - F_{j-1})}{(k+1)(x_j - x_{j-1})}
\end{aligned}$$

$$+x^k\left[1-\frac{(x-x_{j-1})F_j+(x_j-x)F_{j-1}}{x_j-x_{j-1}}\right],$$

当 $k=1$ 时,

$$\mathrm{E}(S^{\#\#}\wedge x)=\sum_{i=2}^{j-1}\frac{(x_i+x_{i-1})(F_i-F_{i-1})}{2}+\frac{(x^2-x_{j-1}^2)(F_j-F_{j-1})}{2(x_j-x_{j-1})}$$
$$+x\left[1-\frac{(x-x_{j-1})F_j+(x_j-x)F_{j-1}}{x_j-x_{j-1}}\right].$$

习题

6.56 令损失频率分布服从参数 $r=2,\beta=2$ 的负二项分布, 令损失分布服从 $\alpha=4,\theta=25$ 的 gamma 分布, 对单次损失的普通免赔额为 25 的保单, 求 $F_S(200)$ 和 $\mathrm{E}(S\wedge 200)$. 用离散间距为 5 的舍入法来离散化损失分布, 并用递归公式求总赔付的分布.

6.57 (续习题 6.51) 赔案数服从参数 $\lambda=5$ 的 Poisson 分布, 单次赔案额度服从参数 $\alpha=0.5,\theta=2\ 500$ 的 gamma 分布. 求以下保单条款的保险公司赔付额的均值, 标准差和 90% 分位点, 任何计算方法都可以使用.

(a) 最大总赔付额为 20 000.

(b) 单次赔案的普通免赔额为 100, 最大赔付额为 10 000, 总赔付额没有上限.

(c) 单次赔案的普通免赔额为 100, 赔付额无上限. 总赔付额的普通免赔额 15 000, 分保因子为 0.8, 最大总赔付额为 20 000—— 这相当于总赔付的再保险条款.

6.58 (续习题 6.52) 索赔次数服从参数 $\lambda_1=10,\lambda_2=4$ 的 Poisson − Poisson 分布, 单次赔付额为 5 的概率是 0.75, 为 10 的概率是 0.25. 求在以下情形投保人赔付的期望值, 可以用任何方法计算.

(a) 最大赔付额 400.

(b) 现有分保合约规定保险人将 100% 赔付总损失量在 300 以内的, 以及将赔付超过 300 部分的 20%.

6.9 反演方法

本节讨论的反演法是从某个已知的变换 —— 如待求函数的概率生成函数, 矩母函数或特征函数, 数值地计算概率函数或是相关的其他函数 —— 如净止损保费 (总超额损失净保费) 的方法.

复合分布很自然地适用于这种方法, 因为当损失频率和损失程度部分都已知时, 它们的变换式是复合函数且容易估计. 总损失分布的概率生成函数和特征函数分别为

$$P_S(z)=P_N[P_X(z)]$$

和

$$\varphi_S(z)=\mathrm{E}[\mathrm{e}^{\mathrm{i}Sz}]=P_N[\varphi_X(z)]. \tag{6.28}$$

特征函数始终存在且唯一. 反过来, 对给定的特征函数, 始终存在唯一的分布与之对应. 反演法的目的是为了由特征函数 (6.28) 式来数值地计算其分布.

值得一提的是在应用概率论的其他领域, 最近出现了许多关于如何由联合 Laplace-Stieltjes 变换数值地计算分布函数的研究. 这些技术可以应用在本章研究的复合分布的求解问题上, 但不在这里进一步讨论了. 文献 [2] 第 257 页到第 323 页的内容对这些方法进行了很好的概述.

6.9.1 快速傅里叶变换

快速傅里叶变换 (FFT) 是通过特征函数的逆函数来得到离散随机变量密度的一种算法. FFT 源于信号处理领域. Bertram[14] 最先用它对特征函数求逆得到复合分布函数. Robertson[111] 则详细说明了如何将该方法应用于总损失计算.

定义 6.26 *对任何连续函数 $f(x)$, 定义如下的傅里叶变换映射*

$$\widetilde{f}(z)=\int_{-\infty}^{\infty} f(x)\mathrm{e}^{\mathrm{i}zx}\mathrm{d}x. \tag{6.29}$$

其中的原始函数可以由它的傅里叶变换表示为

$$f(x)=\frac{1}{2\pi}\int_{-\infty}^{\infty} \widetilde{f}(z)\mathrm{e}^{-\mathrm{i}zx}\mathrm{d}z.$$

当 $f(x)$ 为概率密度函数时, $\widetilde{f}(z)$ 就是它的特征函数. 在应用中, $f(x)$ 是实值的. 由 (6.29) 式知, $\tilde{f}(z)$ 是复值的. 当 $f(x)$ 为离散 (或复合) 分布的概率函数时, 该定义很容易推广 (见 Fisz[38]).

定义 6.27 *记 f_x 为定义在所有整数值 x 上的周期函数, 周期为 n(即对所有的 x, $f_{x+n}=f_x$). 对向量 $(f_0,f_1,\cdots,f_{n-1})$, 离散傅里叶变换 (discrete Fourier transform) 是如下定义的映射 $\widetilde{f}_x, x=\cdots,-1,0,1,\cdots$*

$$\widetilde{f}_k=\sum_{j=0}^{n-1} f_j \exp\left(\frac{2\pi\mathrm{i}}{n}jk\right), \quad k=\cdots,-1,0,1,\cdots. \tag{6.30}$$

此映射是双射, 另外, $\widetilde{f}_k$ 同样是周期为 n 的周期函数. 该映射的逆映射为

$$f_j=\frac{1}{n}\sum_{k=0}^{n-1} \widetilde{f}_k \exp\left(-\frac{2\pi\mathrm{i}}{n}kj\right), \quad j=\cdots,-1,0,1,\cdots. \tag{6.31}$$

由这个逆映射可以恢复原始函数.

由 f 和 $\tilde{f}$ 的周期特性可以想到离散傅里叶变换应该是从 n 个点映到 n 个点的双射. 由 (6.30) 式, 要得到 $\tilde{f}_k$ 的 n 个值, 计算的项数显然为 n^2 阶的, 即 $O(n^2)$.

快速傅里叶变换 (Fast Fourier Transform, FFT) 是将所需计算次数降为 $O(n \ln_2 n)$ 阶的一种算法. 当 n 比较大时, 可以使得计算次数显著减少. 这个算法利用

了离散傅里叶变换的一个性质：长度为 n 的离散傅里叶变换可以写成两个长度为 $n/2$ 的离散变换之和, 第一个变换只包含离散傅里叶变换的偶数点, 第二个变换只包含奇数点.

$$\begin{aligned}\widetilde{f}_k &= \sum_{j=0}^{n-1} f_j \exp\left(\frac{2\pi \mathrm{i}}{n} jk\right) \\ &= \sum_{j=0}^{n/2-1} f_{2j} \exp\left(\frac{2\pi \mathrm{i}}{n} 2jk\right) + \sum_{j=0}^{n/2-1} f_{2j+1} \exp\left[\frac{2\pi \mathrm{i}}{n}(2j+1)k\right] \\ &= \sum_{j=0}^{m-1} f_{2j} \exp\left(\frac{2\pi \mathrm{i}}{m} jk\right) + \exp\left(\frac{2\pi \mathrm{i}}{n} k\right) \sum_{j=0}^{m-1} f_{2j+1} \exp\left(\frac{2\pi \mathrm{i}}{m} jk\right),\end{aligned}$$

若 $m = n/2$, 则有

$$\widetilde{f}_k = \widetilde{f}_k^a + \exp\left(\frac{2\pi \mathrm{i}}{n} k\right) \widetilde{f}_k^b. \tag{6.32}$$

上式右边两项可以分别写成两个长度为 $m/2$ 的变换之和, 如此进行下去. 为了使长度 $n/2, m/2, \cdots$ 均为整数, FFT 算法应由一长度为 $n = 2^r$ 的向量开始, 然后将每一步所得的结果写成长度减半的变换之和, 最终在 r 次分解之后得到长度为 1 的变换. 若已知了长度为 1 的变换, 利用 (6.32) 式的简单叠加就可以逐步得到长度为 $2, 2^2, 2^3, \cdots, 2^r$ 的变换. 此方法详见 Press et al. [107].

在应用中, 我们是在损失程度分布离散化之后, 使用 FFT 来对特征函数求逆, 其步骤如下.

(1) 用某些前面几节中介绍的方法离散化损失分布, 得到离散的损失分布为

$$f_X(0), f_X(1), \cdots, f_X(n-1),$$

其中 $n = 2^r$, r 为整数, n 为估计总索赔额分布 $f_S(x)$ 所需要的离散点的数量.

(2) 对这个向量值由 FFT 得到离散化分布的特征函数 $\varphi_X(z)$, 其结果也是一个 $n = 2^r$ 维的向量.

(3) 用索赔频率分布的概率生成函数变换对这个向量进行变换, 得到特征函数 $\varphi_S(z) = P_N[\varphi_X(z)]$, 即总索赔额分布的离散傅里叶变换, 也是一个 $n = 2^r$ 维的向量.

(4) 应用快速傅里叶逆变换 (Inverse Fast Fourier Transform, IFFT)—— 除了有一处符号变化以及结果要除以 n 外 [见 (6.31) 式], 其他计算都和 FFT 一样. 这就是基于离散化损失程度模型给出了一个能够代表总赔付额精确分布的 $n = 2^r$ 维向量.

FFT 的过程需要离散化损失分布. 当损失分布的点数少于 $n = 2^r$, 必须在损失分布向量上添一些 0 将长度补齐到 n.

当损失分布在 $x = n$ 以上的部分还有概率时, 比如在第 4 章中讨论过的大多数分布, 丢弃 n 之后的右尾部概率会导致最终结论的微小误差. 因为函数和它的变换式都被假设成以 n 为周期, 而事实并非如此. 有学者建议将剩下的所有概率都放在最终点 $x = n$ 上, 从而使得总概率加起来等于 1, 进而可以在 FFT 算法中对损失分布应用周期性假设并确保最终的总损失概率之和等于 1. 但是, 必须选取足够大的 n 值, 使得几乎所有的总损失概率都分布在第 n 个点之前. 下面的例题提供了一个极端的例子.

例 6.28 **设随机变量 X 等于 1, 2, 3 的概率分别为 0.5, 0.4, 0.1, 并假设索赔数服从参数 $\lambda = 3$ 的 Poisson 分布. 使用 FFT, 当分别取 n=8 和 n=4 096, 求 S 的分布.**

解 在两种情形下, X 的概率分布都由 0 开始 (由于 S 在 0 点有概率, X 的初始表达式也必须有给出零点的概率), 并有 4 个或 4 092 个 0 在尾端. 使用 FFT 和 IFFT 的结果在表 6-17 中给出. n=8 时, 8 个概率之和为 1; n=4 096 时, 概率和也为 1, 但此处没有足够的空间将这些概率值全部列出. 这个问题很容易用递归公式进行检验, 若检验 n=4 096 的各个项, 可以发现计算的每一项结果在给出的 5 位小数下都是精确的. 另一方面, n=8 时, FFT 给出的值显然是误差较大的. 如果想要推广该方法, 就应当将剩余的概率分配一些到较小的 S 值上.

表 6-17 由 FFT 和 IFFT 复合得到的总概率

s	$n = 8$ $f_S(s)$	$n = 4\ 096$ $f_S(s)$
0	0.112 27	0.049 79
1	0.118 21	0.074 68
2	0.144 70	0.115 75
3	0.151 00	0.132 56
4	0.147 27	0.135 97
5	0.131 94	0.125 25
6	0.109 41	0.105 58
7	0.085 18	0.083 05

由于很多计算机软件包中都含有 FFT 和 IFFT 算法, 且程序代码较短、容易编写甚至可以直接得到 (例如, 文献 [107] 第 411 页到第 412 页), 这里就不再给出关于这种算法的更多技术细节. 若读者想了解更多的细节, 可以从众多关于 FFT 的书籍中任选一本阅读. 令计算速度由 $O(n^2)$ 阶变为 $O(n \ln_2 n)$ 阶的技术细节涉及离散傅里叶变换的具体性质. Robertson[111] 对如何应用 FFT 计算总赔付额分布给出了很好的阐释. □

6.9.2 直接数值反演

对 Poisson, 二项和负二项损失频率分布以及连续损失分布, Heckman 和 Meyers 已经在文献 [51] 中使用近似积分的方法得到了特征函数 (6.28) 式的反演方法, 这种方法很容易推广到其他索赔频率分布的情形.

在这种方法中, 损失分布由分段线性分布函数代替, 并使用了单次损失最大赔付额, 进而累积分布函数在个体单次损失最大值处跳到 1. 损失程度随机变量的分布范围被划分为长度可能不等的区间. 剩下的步骤和 FFT 方法的步骤相对应. 考虑损失程度的累积分布函数 $F_X(x), 0 \leqslant x < \infty$, 令 $0 = x_0 < x_1 < \cdots < x_n$ 是任意选择的损失值, 损失额在区间 $(x_{k-1}, x_k]$ 上的概率记为 $f_k = F_X(x_k) - F_X(x_{k-1})$. 在这些区间上使用均匀密度 d_k, 得到近似的密度函数 $f^*(x) = d_k = f_k/(x_k - x_{k-1}), x_{k-1} < x \leqslant x_k$, 所有的剩余概率 $f_{n+1} = 1 - F_X(x_n)$ 作为尖柱放在 x_n 上. 如此对概率密度函数进行近似处理将使得特征函数的公式较为简单, 但在直接反演法中这并不是必须的. 损失程度近似分布的特征函数为

$$\begin{aligned}\varphi_X(z) &= \int_0^{\infty} \mathrm{e}^{\mathrm{i}zx} \mathrm{d}F_X(x) \\ &= \sum_{k=1}^{n} \int_{x_{k-1}}^{x_k} \mathrm{d}_k \mathrm{e}^{\mathrm{i}zx} \mathrm{d}x + f_{n+1} \mathrm{e}^{\mathrm{i}zx_n} \\ &= \sum_{k=1}^{n} \mathrm{d}_k \frac{\mathrm{e}^{\mathrm{i}zx_k} - \mathrm{e}^{\mathrm{i}zx_{k-1}}}{\mathrm{i}z} + f_{n+1} \mathrm{e}^{\mathrm{i}zx_n}.\end{aligned}$$

由欧拉公式

$$\mathrm{e}^{\mathrm{i}\theta} = \cos(\theta) + \mathrm{i}\sin(\theta).$$

特征函数可以分为实部和虚部. 实部为

$$a(z) = \mathrm{Re}[\varphi_X(z)] = \frac{1}{z} \sum_{k=1}^{n} d_k[\sin(zx_k) - \sin(zx_{k-1})] + f_{n+1}\cos(zx_n),$$

虚部为

$$b(z) = \mathrm{Im}[\varphi_X(z)] = \frac{1}{z} \sum_{k=1}^{n} d_k[\cos(zx_{k-1}) - \cos(zx_k)] + f_{n+1}\sin(zx_n).$$

进而可得到总损失 (6.28) 式的特征函数

$$\varphi_S(z) = P_N[\varphi_X(z)] = P_N[a(z) + \mathrm{i}b(z)],$$

由于它是复数, 故可以写成

$$\varphi_S(z) = r(z)\mathrm{e}^{\mathrm{i}\theta(z)}.$$

由此可以得到总赔付额的分布为

$$F_S(x) = \frac{1}{2} + \frac{1}{\pi}\int_0^\infty \frac{r(z/\sigma)}{z}\sin\left(\frac{zx}{\sigma} - \frac{\theta z}{\sigma}\right)\mathrm{d}z, \tag{6.33}$$

其中 σ 是总损失分布的标准差. 对任意值 x, 可采用近似积分技术来计算 (6.33) 式, 读者可以参考 Heckman and Meyers[51] 以了解其细节. 由 (6.33) 式, 同样可以根据总损失分布得到净止损 (超额净) 保费如下

$$\begin{aligned} P(d) &= \mathrm{E}[(S-d)_+] = \int_d^\infty (s-d)\mathrm{d}F_S(s) \\ &= \frac{\sigma}{\pi}\int_0^\infty \frac{r(z/\sigma)}{z^2}\left[\cos\left(\frac{\theta z}{\sigma}\right) - \cos\left(\frac{zd}{\sigma} - \frac{\theta z}{\sigma}\right)\right]\mathrm{d}z + \mu - \frac{d}{2}, \end{aligned} \tag{6.34}$$

其中 μ 是总损失分布的均值, d 是免赔额.

尽管 (6.33) 式仅给出了分布的值, (6.34) 式也仅给出了净保费的值, 但它们计算起来相当迅速. 近似的误差取决于数值积分方法的间距是可以控制的.

习题

6.59 用反演法重新解答习题 6.51 和 6.52.

6.10 不同方法的比较

递归法有一些明显的优势. 相比于直接卷积法, 计算整个分布中 n 个点所需的时间由 $O(n^3)$ 减少到 $O(n^2)$ 阶. 而且, 若损失分布本身是离散的 (算术的), 得到的就是精确值. 唯一可能产生误差的步骤是损失分布的离散化. 除了二项分布, 计算可以保证是数值稳定的. 这种方法容易编程, 只须几行程序代码. 尽管如此, 它也有一些缺点. 递归法仅适用于 4.6 节中讨论的索赔频率分布类, 若不是 $(a,b,0)$ 或 $(a,b,1)$ 分布类时, 需要对公式进行修正或给出新的递归式. 近来, 在精算和统计文献中已经出现了多种其他递归式.

FFT 法容易使用, 因为它使用的是许多软件包里都采用的标准程序. 当 n 很大时, 它比递归法计算速度快一些, 因为它需要的计算次数是 $n\ln_2 n$ 阶的, 而不是 n^2 阶. 但是, 若损失分布于数目为常数 (且数目不是很大) 的点上, 则递归法需要的计算次数更少, 因为 (6.14) 式中的和式最多只有 m 项, 所需的计算次数是 n 阶的, 而不是没有损失程度上限情形下的 n^2 阶. FFT 法可以推广到损失分布取负值的情形. 与递归法类似, 它产生的是整个分布.

可以证明, 直接反演法用来计算总分布或免赔额为 d 时的净止损 (超损净) 保费的单个值是非常快的, 但这需要在计算机程序上花更多的功夫. 该方法已经在 Heckman and Meyers[51] 中应用于 $(a,b,0)$ 类的损失频率模型, 并且只要概率生成函数相对较为简单, 该文献涉及的计算机代码可以推广到任何分布. 当索赔次数的期望值较大时, 这种方法的计算速度比递归法要快得多. 在 Poisson 索赔频率模型下, 计算速度和 λ 值的大小无关. 当然, 该方法不仅编程较为麻烦, 还包含了近似积分, 从而其误差取决于近似积分的方法和积分区间的大小.

FFT 法和反演法通过使用变换式有效地解决了卷积的问题. 例如, 假设某再保险合约承保了三个团体的总损失, 每个团体都有各自的损失频率和损失分布. 若 S_i, $i=1,2,3$ 是三个团体分别的总损失, 则总损失 $S=S_1+S_2+S_3$ 的特征函数为 $\varphi_S(z)=\varphi_{S_1}(z)\varphi_{S_2}(z)\varphi_{S_3}(z)$, 因此, 剩下的工作仅仅是反演一些乘法运算. 但递归法就不能如此轻易地处理卷积.

如果损失分布是离散的或不规则的, 如损失集中在某个舍入值 (如 1 000 000) 上的情形, 运用 Heckman-Meyers 方法存在一些技术困难. 在设定点 $(x_1,x_2,\cdots,x_n)$ 时, 需要在损失分布函数的每个跳跃点定义一个包含该跳跃点的小区间.

将随机模拟法留到最后讨论是因为它和其他方法有很大不同. 考虑到不少读者对随机模拟还不甚熟悉, 第 17 章对此进行了简要介绍. 这种方法的优点十分明显 —— 只要能够详细地给出模型, 就可以通过模拟得到总损失分布. 也许编程会花一些时间, 但编程语言却采用了直截了当的方式. 如今, 计算机能在合理的时间内执行模拟过程, 虽然有的模拟过程需要较多时间, 但有大量的解析方法出现, 使得这些模拟过程现在也不成问题了. 另一方面, 要写一个普遍适用的模拟程序是很困难的. 相反, 很可能需要为每一个新遇到的问题编写新的程序. 随机模拟法也许是处理那些其他方法无法解决的问题的最好方法, 因为在没有其他选择的时候, 在随机模拟上下功夫显然是十分明智的选择.

当索赔频率很低时 (此时递归法胜过其他方法), 随机模拟法存在另一个缺点. 例如, 考虑一份个体超额损失再保险, 它赔付个体损失额超过 1 000 000 的部分. 损失超过免赔额的可能性约为 1/100, 但这种损失一旦发生, 其尾部非常长 (如 α 值很小的 Pareto 分布). 模拟过程不得不丢弃生成的损失额中的 99%, 由于需要生成大量超过免赔额的损失 (因为损失额的大方差), 因而也许会要很长的时间才能得到可靠的结果. 一种可能的解决方案是, 给定赔付已经发生的条件下, 用损失变量的条件分布来计算.

没有哪种方法是对所有问题都明显优于其他方法的. 与其他方法相比, 每种方法都有自己的优点和缺点 —— 实际上我们面临的是可选方案过多的窘境. 25 年前, 精算学家们还在怀疑是否存在求总分布的有效方法, 而如今我们已可以从多种方法中挑选.

6.11 个体风险模型

6.11.1 参数的近似

个体风险模型 (individual risk model) 表示数量确定的若干独立随机变量 (但不需要同分布) 之和:

$$S = X_1 + X_2 + \cdots + X_n.$$

通常可以将其看成 n 个保单合同的损失之和, 例如, 某个保单组由 n 个投保人组成.

个体风险模型最初是为一类寿险保单设计的, 在这类保单中, 第 j 个投保人在一年内身故的概率为 q_j, 该投保人的身故保险金为 b_j. 此时, 对第 j 份保单, 保险人的损失分布为

$$f_{X_j}(x) = \begin{cases} 1 - q_j, & x = 0, \\ q_j, & x = b_j. \end{cases}$$

由于假设了 X_j 的独立性, 总损失的均值和方差为

$$\mathrm{E}(S) = \sum_{j=1}^{n} b_j q_j,$$

$$\mathrm{Var}(S) = \sum_{j=1}^{n} b_j^2 q_j (1 - q_j).$$

总损失的概率生成函数为

$$P_S(z) = \prod_{j=1}^{n} (1 - q_j + q_j z^{b_j}). \qquad (6.35)$$

在这种特殊情形下, 所有的风险相同, 即 $q_j = q$ 且 $b_j = 1$, 则概率生成函数可以化简为

$$P_S(z) = [1 + q(z - 1)]^n,$$

S 服从二项分布.

个体风险模型可由如下方法生成: 令 $X_j = I_j B_j$, 其中 $I_1, \cdots, I_n, B_1, \cdots, B_n$ 独立, 随机变量 I_j 为示性变量, 取 1 的概率为 q_j, 取 0 的概率为 $1 - q_j$, 该变量用来指示第 j 个保单是否产生了赔付; 随机变量 B_j 可以服从任何分布, 它表示第 j 份保单在已知赔付发生的条件下的赔付额. 在寿险案例中, B_j 为退化的, 只有 b_j 一个取值. 令 $\mu_j = \mathrm{E}(B_j), \sigma_j^2 = \mathrm{Var}(B_j)$, 则

$$\mathrm{E}(S) = \sum_{j=1}^{n} q_j \mu_j, \qquad (6.36)$$

$$\mathrm{Var}(S)=\sum_{j=1}^{n}[q_j\sigma_j^2+q_j(1-q_j)\mu_j^2]. \tag{6.37}$$

习题 6.60 要求读者验证上述公式. 下面这个例题是该情形中较为简单的一个例子.

例 6.29 考虑一组包含意外死亡赔付的寿险合同, 假设所有成员在未来一年内死亡的概率都是 0.01, 30%的死亡属于意外死亡. 对其中 50 位投保人, 正常死亡保险金是 50 000, 意外死亡保险金是 100 000; 剩下 25 位投保人的保险金分别为 75 000 和 150 000. 给出个体风险模型并求它的均值和方差.

解 对所有的 75 位投保人, $q_j=0.01$. 对其中 50 位投保人, B_j 等于 50 000 的概率为 0.7, 等于 100 000 的概率为 0.3, $\mu_j=65\ 000, \sigma_j^2=525\ 000\ 000$; 对剩下的 25 位投保人, B_j 等于 75 000 的概率为 0.7, 等于 150 000 的概率为 0.3, $\mu_j=97\ 500, \sigma_j^2=1\ 181\ 250\ 000$. 此时

$$\mathrm{E}(S)=50(0.01)(65\ 000)+25(0.01)(97\ 500)=56\ 875,$$

$$\begin{aligned}\mathrm{Var}(S)=&\ 50(0.01)(525\ 000\ 000)+50(0.01)(0.99)(65\ 000)^2\\&+25(0.01)(1\ 181\ 250\ 000)+25(0.01)(0.99)(97\ 500)^2=5\ 001\ 984\ 375.\end{aligned}$$

当风险不同时, 由概率生成函数 (6.35) 式定义的概率既可以精确计算也可以近似计算. 正态分布, gamma 分布, 对数正态分布或任何其他分布都可以用来近似计算其分布, 通常由匹配前几阶矩的方法完成. 由于正态分布, gamma 分布, 对数正态分布都只有两个参数, 所以匹配均值和方差就足够了. □

例 6.30 (团体寿险) 一小型生产企业为其 14 位永久雇员签订了一份团体寿险合同, 保险公司的精算师选择了一份生命表来反映这个团体的死亡力, 每位被保险雇员的保险金是将工资按照四舍五入原则取最接近的千元单位. 这个团体的数据在表 6-18 中给出.

如果保险公司在净保费上增加 45%的安全附加, 在一给定年份内, 保险公司在该保单组亏损的概率为多少? 使用正态和对数正态近似.

解 该团体总损失的均值和方差为

$$\mathrm{E}(S)=\sum_{j=1}^{14}b_jq_j=2\ 054.41,$$

$$\mathrm{Var}(S)=\sum_{j=1}^{14}b_j^2q_j(1-q_j)=1.025\ 34\times 10^8.$$

收取的保费为 $1.45\times 2\ 054.41=2\ 978.89$, 正态近似 (单位为 1 000), 均值为 2.054 41, 方差为 102.534, 则损失的概率为

$$\Pr(S > 2.978\ 89) = \Pr\left[Z > \frac{2.978\ 89 - 2.054\ 41}{(102.534)^{1/2}}\right] = \Pr(Z > 0.091\ 3) = 0.46.$$

表 6-18 例 6.30 中投保人的数据

投保人 j	年龄 (年)	性别	赔付额 b_j	死亡率 q_j
1	20	男	15 000	0.001 49
2	23	男	16 000	0.001 42
3	27	男	20 000	0.001 28
4	30	男	28 000	0.001 22
5	31	男	31 000	0.001 23
6	46	男	18 000	0.003 53
7	47	男	26 000	0.003 94
8	49	男	24 000	0.004 84
9	64	男	60 000	0.021 82
10	17	女	14 000	0.000 50
11	22	女	17 000	0.000 50
12	26	女	19 000	0.000 54
13	37	女	30 000	0.001 03
14	55	女	55 000	0.004 79
总计			373 000	

对数正态近似 (与例 6.5 类似),

$$\mu + \frac{1}{2}\sigma^2 = \ln 2.054\ 41 = 0.719\ 989,$$
$$2\mu + 2\sigma^2 = \ln(102.534 + 2.054\ 41^2) = 4.670\ 533.$$

由此得到 $\mu = -0.895\ 289, \sigma^2 = 3.230\ 555$, 那么

$$\Pr(S > 2.978\ 89) = 1 - \Phi\left[\frac{\ln 2.978\ 89 + 0.895\ 289}{(3.230\ 555)^{1/2}}\right] = 1 - \Phi(1.105) = 0.13. \quad \square$$

6.11.2 节将给出几种在保险金固定的情形下获得 S 精确分布的方法.

6.11.2 总分布的精确计算

直接计算

总损失的概率函数由

$$f_S(x) = f_{X_1} * f_{X_2} * \cdots * f_{X_n}(x), \tag{6.38}$$

给出, 其中

$$f_{X_j}(x) = \begin{cases} p_j = 1 - q_j, & x = 0, \\ q_j, & x = b_j. \end{cases}$$

(6.38) 式的密度可以由部分和 $S_j = S_{j-1} + X_j, j = 2, 3, \cdots, n$ 递归地计算, 其初值 $S_1 = X_1$. 则

$$f_{S_j}(x) = \begin{cases} f_{S_{j-1}}(x) f_{X_j}(0), & x < b_j, \\ f_{S_{j-1}}(x) f_{X_j}(0) + f_{S_{j-1}}(x-b_j) f_{X_j}(b_j), & x \geqslant b_j, \end{cases}$$
$$= \begin{cases} p_j f_{S_{j-1}}(x), & x < b_j, \\ p_j f_{S_{j-1}}(x) + q_j f_{S_{j-1}}(x-b_j), & x \geqslant b_j. \end{cases}$$

如果想计算某个 r 以内的总赔付额分布, 所需的计算时间是以乘法计量的为 nr 阶. 如果 r 和 n 都很大 (如 $r = 10\ 000, n = 1\ 000$), 计算次数将大得惊人. □

例 6.31 (续例 6.30) 使用直接法计算 S 的概率函数和例 6.30 里所求的概率.

解

$$\begin{aligned} f_{S_1}(0) &= 0.998\ 51, \\ f_{S_1}(15) &= 0.001\ 49, \\ f_{S_2}(0) &= p_2 f_{S_1}(0) = 0.997\ 092\ 12, \\ f_{S_2}(15) &= p_2 f_{S_1}(15) = 0.001\ 487\ 88, \\ f_{S_2}(16) &= p_2 f_{S_1}(16) + q_2 f_{S_1}(0) = 0.001\ 417\ 88, \\ f_{S_2}(31) &= p_2 f_{S_1}(31) + q_2 f_{S_1}(15) = 0.000\ 002\ 115\ 8. \end{aligned}$$

对 $x = 0, \cdots, 79$, 分布函数 $F_S(x)$ 的值如表 6-19 所示. 从表 6-19 可以看出, 超过 2 978.89 的概率为 0.047, 由此可见例 6.30 中两种方法的近似程度都很差. □

当 n 不是很大时, 这种方法是可行的, 但对较大的群体就需要另一种方法.

递归计算

下面的方法可以基于 De Pril[26] 的方法, 递归地计算分布函数. 首先需要根据保单大小和索赔概率将保单组合划分为多个子组. 设 n_{ij} 为保险金为 i (其中 $i = 1, 2, \cdots, r$)[①], 索赔概率为 q_j(其中 $j = 1, 2, \cdots, m$) 的保单数目, 则总赔付额的概率生成函数可以写成

$$P_S(z) = \prod_{i=1}^{r} \prod_{j=1}^{m} (1 - q_j + q_j z^i)^{n_{ij}}.$$

概率生成函数的对数为

$$\ln P_S(z) = \sum_{i=1}^{r} \sum_{j=1}^{m} n_{ij} \ln(1 - q_j + q_j z^i). \tag{6.39}$$

① 正如对损失分布的离散化一样也需要对赔付额进行算术化处理. 但是, 货币单位可以不是 1, 比如, $i = 1, 2, \cdots$ 表示的赔付额可以为 $5\ 000, 10\ 000, \cdots$

对 (6.39) 式求导得到

$$P'_S(z) = P_S(z)\left[\sum_{i=1}^{r}\sum_{j=1}^{m} iq_j n_{ij} z^{i-1}(1-q_j+q_j z^i)^{-1}\right]. \tag{6.40}$$

表 6-19　例 6.31 的累积概率

x	$F_S(x)$	x	$F_S(x)$	x	$F_S(x)$	x	$F_S(x)$
0	0.952 739 05	20	0.961 579 69	40	0.973 350 98	60	0.999 330 62
1	0.952 739 05	21	0.961 579 69	41	0.973 358 92	61	0.999 331 87
2	0.952 739 05	22	0.961 579 69	42	0.973 381 28	62	0.999 331 91
3	0.952 739 05	23	0.961 579 69	43	0.973 387 40	63	0.999 331 93
4	0.952 739 05	24	0.966 213 37	44	0.973 408 84	64	0.999 331 98
5	0.952 739 05	25	0.966 213 37	45	0.973 413 51	65	0.999 332 02
6	0.952 739 05	26	0.969 982 01	46	0.973 425 61	66	0.999 332 06
7	0.952 739 05	27	0.969 982 01	47	0.973 428 40	67	0.999 332 09
8	0.952 739 05	28	0.971 145 77	48	0.973 433 97	68	0.999 332 17
9	0.952 739 05	29	0.971 146 48	49	0.973 438 66	69	0.999 334 50
10	0.952 739 05	30	0.972 129 50	50	0.973 458 89	70	0.999 341 41
11	0.952 739 05	31	0.973 305 07	51	0.973 460 40	71	0.999 347 96
12	0.952 739 05	32	0.973 307 47	52	0.973 466 06	72	0.999 350 31
13	0.952 739 05	33	0.973 313 44	53	0.973 466 08	73	0.999 366 59
14	0.953 215 66	34	0.973 319 62	54	0.973 475 47	74	0.999 379 73
15	0.954 637 36	35	0.973 323 86	55	0.978 066 78	75	0.999 417 35
16	0.955 992 17	36	0.973 325 85	56	0.978 070 68	76	0.999 447 59
17	0.956 468 78	37	0.973 328 29	57	0.978 075 36	77	0.999 458 23
18	0.959 843 86	38	0.973 334 93	58	0.978 076 60	78	0.999 533 55
19	0.960 358 62	39	0.973 342 51	59	0.978 078 08	79	0.999 567 34

令 (6.40) 式中的 z=1, 可得到总赔付额分布的均值, 形式上

$$\mathrm{E}(S) = P'_S(1) = \sum_{i=1}^{r}\sum_{j=1}^{m} iq_j n_{ij}.$$

现在, (6.40) 式可以改写为

$$\begin{aligned} zP'_S(z) &= P_S(z)\left[\sum_{i=1}^{r}\sum_{j=1}^{m} in_{ij}\left(\frac{q_j}{1-q_j}z^i\right)\left(1+\frac{q_j}{1-q_j}z^i\right)^{-1}\right] \\ &= P_S(z)\left[\sum_{i=1}^{r}\sum_{j=1}^{m} in_{ij}\sum_{k=1}^{\infty}(-1)^{k-1}\left(\frac{q_j}{1-q_j}\right)^k z^{ik}\right], \end{aligned} \tag{6.41}$$

对 $|z| < \min_{i,j}[q_j^{-1}(1-q_j)]^{1/i}$ 成立. (6.41) 式右边第二项可以改写为

$$\sum_{i=1}^{r}\sum_{k=1}^{\infty} h(i,k)z^{ik},$$

其中

$$h(i,k) = i(-1)^{k-1}\sum_{j=1}^{m} n_{ij}\left(\frac{q_j}{1-q_j}\right)^k. \tag{6.42}$$

因此, (6.41) 式可以改写为

$$zP_S'(z) = P_S(z)\left[\sum_{i=1}^{r}\sum_{k=1}^{\infty} h(i,k)z^{ik}\right]. \tag{6.43}$$

(6.43) 左边 z^x 的系数是 $xf_S(x)$, 其中 $f_S(x)$ 是 $P_S(z)$ 中 z^x 的系数; (6.43) 式右边为一个卷积, 因此 z^x 的系数为

$$\sum_{ik\leqslant x} h(i,k)f_S(x-ik). \tag{6.44}$$

(6.44) 式更简单的写法为

$$\sum_{i=1}^{x}\sum_{k=1}^{\lfloor x/i\rfloor} h(i,k)f_S(x-ik),$$

其中 $\lfloor\cdot\rfloor$ 为最大整数函数的记号, 即小于或等于自变量的最大整数. 最后, 由于当 $i > x$ 时 $h(i,k) = 0$, 比较 (6.43) 式两边 z^x 的系数并除以 x 得到

$$f_S(x) = \frac{1}{x}\sum_{i=1}^{x\wedge r}\sum_{k=1}^{\lfloor x/i\rfloor} h(i,k)f_S(x-ik), \quad x\geqslant 1. \tag{6.45}$$

则有

$$f_S(0) = P_S(0) = \prod_{i=1}^{r}\prod_{j=1}^{m}(1-q_j)^{n_{ij}}, \tag{6.46}$$

并由 (6.45) 式, 得

$$\begin{aligned}
f_S(1) &= h(1,1)f_S(0),\\
f_S(2) &= \frac{1}{2}\{h(1,1)f_S(1) + [h(1,2)+h(2,1)]f_S(0)\},\\
&\ \ \vdots
\end{aligned}$$

$\{f_S(x); x = 1,2,\cdots\}$ 的概率可以用 (6.45) 式进行递归计算, 初值为 (6.46) 式.

由 (6.42) 式可见, $h(i,k)$ 为 $q_j/(1-q_j), j=1,2,\cdots,m$ 的 k 次幂的加权求和. 当 q_j 接近 0 时, $[q_j/(1-q_j)]^k$ 很小, 因此 $h(i,k)$ 的值随着 k 的增加快速递减. 这表明可以将 (6.45) 式的内层和号的项数限制在较小的范围内, 此举不会对计算精度造成明显的损失, 但却在很大程度上提高了计算速度.

如果限定 k 最多为 K 项, 记

$$f_S^{(K)}(x) = \frac{1}{x}\sum_{i=1}^{x\wedge r}\sum_{k=1}^{K\wedge\lfloor x/i\rfloor} h(i,k) f_S^{(K)}(x-ik) \tag{6.47}$$

为 (6.45) 式中将求和式限制为 K 项的近似值. De Pril[27] 证明了以下结论：若 $q_j<\frac{1}{2}, j=1,2,\cdots m$, 则

$$\sum_{x=0}^{M}|f_S(x)-f_S^{(K)}(x)| < \mathrm{e}^{\delta(K)}-1, \tag{6.48}$$

其中

$$\delta(K)=\frac{1}{K+1}\sum_{i=1}^{r}\sum_{j=1}^{m} n_{ij}\frac{1-q_j}{1-2q_j}\left(\frac{q_j}{1-q_j}\right)^{K+1}, \tag{6.49}$$

$M=\sum\limits_{i=1}^{r}\sum\limits_{j=1}^{m} i n_{ij}$ 为总索赔额的最大可能值.

对任意值 K 很容易计算 $\delta(K)$ 的值. 等式 (6.48) 给出了总索赔额在整个分布范围内绝对误差之和的上限, 并可以在 (6.47) 式中使用项数限制保证结果的精确度.

例 6.32 (续例 6.30) *用递归法求总损失的精确分布.*

解 表 6-20 给出了 q_j 值和 n_{ij} 所构成矩阵中非零行向量的值.

表 6-20 例 6.32 中 n_{ij} 的值

j	$1000q_j$	i													
		14	15	16	17	18	19	20	24	26	28	30	31	55	60
1	0.50	1	0	0	1	0	0	0	0	0	0	0	0	0	0
2	0.54	0	0	0	0	0	1	0	0	0	0	0	0	0	0
3	1.03	0	0	0	0	0	0	0	0	0	0	1	0	0	0
4	1.22	0	0	0	0	0	0	0	0	0	1	0	0	0	0
5	1.23	0	0	0	0	0	0	0	0	0	0	0	1	0	0
6	1.28	0	0	0	0	0	0	1	0	0	0	0	0	0	0
7	1.42	0	0	1	0	0	0	0	0	0	0	0	0	0	0
8	1.49	0	1	0	0	0	0	0	0	0	0	0	0	0	0
9	3.53	0	0	0	0	1	0	0	0	0	0	0	0	0	0
10	3.94	0	0	0	0	0	0	0	0	1	0	0	0	0	0
11	4.79	0	0	0	0	0	0	0	0	0	0	0	0	1	0
12	4.84	0	0	0	0	0	0	0	1	0	0	0	0	0	0
13	21.82	0	0	0	0	0	0	0	0	0	0	0	0	0	1

利用 (6.49) 式得到 $\delta(1)=5.947\times10^{-4}$, $\delta(2)=3.900\times10^{-6}$, $\delta(3)=6.369\times10^{-8}$ 和 $\delta(4)=1.131\times10^{-9}$, 因此, $K=4$ 时 (6.47) 式可以确保小数点后大约 8 位的精确性. 表 6-21 给出了使用 (6.42) 式计算的 $h(i,k)$ 的 (非零) 值.

表 6-21 例 6.32 中 $h(i,k)$ 的值

i	$h(i,k)$			
	$k=1$	$k=2$	$k=3$	$k=4$
14	$7.003\ 501\ 8\times10^{-3}$	$-3.503\ 502\ 5\times10^{-6}$	$1.752\ 627\ 6\times10^{-9}$	$-8.767\ 521\ 8\times10^{-13}$
15	$2.238\ 335\ 1\times10^{-2}$	$-3.340\ 096\ 2\times10^{-5}$	$4.984\ 169\ 4\times10^{-8}$	$-7.437\ 494\ 4\times10^{-11}$
16	$2.275\ 230\ 9\times10^{-2}$	$-3.235\ 422\ 1\times10^{-5}$	$4.600\ 832\ 5\times10^{-8}$	$-6.542\ 472\ 3\times10^{-11}$
17	$8.504\ 252\ 2\times10^{-3}$	$-4.254\ 253\ 1\times10^{-6}$	$2.128\ 190\ 7\times10^{-9}$	$-1.064\ 627\ 7\times10^{-12}$
18	$6.376\ 509\ 0\times10^{-2}$	$-2.258\ 881\ 6\times10^{-4}$	$8.002\ 099\ 1\times10^{-7}$	$-2.834\ 747\ 7\times10^{-9}$
19	$1.026\ 554\ 3\times10^{-2}$	$-5.546\ 388\ 4\times10^{-6}$	$2.996\ 668\ 0\times10^{-9}$	$-1.619\ 075\ 0\times10^{-12}$
20	$2.563\ 281\ 0\times10^{-2}$	$-3.285\ 204\ 8\times10^{-5}$	$4.210\ 451\ 5\times10^{-8}$	$-5.396\ 285\ 1\times10^{-11}$
24	$1.167\ 249\ 5\times10^{-1}$	$-5.676\ 963\ 8\times10^{-4}$	$2.761\ 013\ 9\times10^{-6}$	$-1.342\ 830\ 0\times10^{-8}$
26	$1.028\ 452\ 1\times10^{-1}$	$-4.068\ 129\ 7\times10^{-4}$	$1.609\ 183\ 3\times10^{-6}$	$-6.365\ 261\ 2\times10^{-9}$
28	$3.420\ 172\ 6\times10^{-2}$	$-4.177\ 707\ 4\times10^{-5}$	$5.103\ 028\ 7\times10^{-8}$	$-6.233\ 299\ 6\times10^{-11}$
30	$3.093\ 186\ 0\times10^{-2}$	$-3.189\ 266\ 5\times10^{-5}$	$3.288\ 331\ 5\times10^{-8}$	$-3.390\ 473\ 6\times10^{-11}$
31	$3.817\ 695\ 9\times10^{-2}$	$-4.701\ 548\ 7\times10^{-5}$	$5.790\ 026\ 6\times10^{-8}$	$-7.130\ 503\ 3\times10^{-11}$
55	$2.647\ 180\ 0\times10^{-1}$	$-1.274\ 102\ 2\times10^{-3}$	$6.132\ 323\ 2\times10^{-6}$	$-2.951\ 520\ 7\times10^{-8}$
60	$1.338\ 40\ 40$	$-2.985\ 542\ 0\times10^{-2}$	$6.659\ 768\ 9\times10^{-4}$	$-1.485\ 576\ 8\times10^{-5}$

利用 $K=4$ 时的 (6.47) 式和 (6.46) 式计算得到的 $f_S(x)$ 值和相应的累积分布函数 $F_S(x)$ 值见表 6-22.

表 6-22 例 6.32 的总概率

x	$f_S(x)$	$F_S(x)$	x	$f_S(x)$	$F_S(x)$
0	$9.527\ 390\ 5\times10^{-1}$	0.952 739 05	38	$6.643\ 643\ 9\times10^{-6}$	0.973 334 93
14	$4.766\ 078\ 3\times10^{-4}$	0.953 215 66	39	$7.574\ 225\ 3\times10^{-6}$	0.973 342 51
15	$1.421\ 699\ 5\times10^{-3}$	0.954 637 36	40	$8.474\ 450\ 8\times10^{-6}$	0.973 350 98
16	$1.354\ 813\ 3\times10^{-3}$	0.955 992 17	41	$7.941\ 654\ 3\times10^{-6}$	0.973 358 92
17	$4.766\ 078\ 3\times10^{-4}$	0.956 468 78	42	$2.235\ 610\ 0\times10^{-5}$	0.973 381 28
18	$3.375\ 082\ 9\times10^{-3}$	0.959 843 86	43	$6.125\ 396\ 1\times10^{-6}$	0.973 387 40
19	$5.147\ 570\ 6\times10^{-4}$	0.960 358 62	44	$2.143\ 544\ 8\times10^{-5}$	0.973 408 84
20	$1.221\ 069\ 0\times10^{-3}$	0.961 579 69	45	$4.672\ 158\ 4\times10^{-6}$	0.973 413 51
24	$4.633\ 684\ 0\times10^{-3}$	0.966 213 37	46	$1.210\ 076\ 9\times10^{-5}$	0.973 425 61
26	$3.768\ 640\ 3\times10^{-3}$	0.969 982 01	47	$2.791\ 514\ 0\times10^{-6}$	0.973 428 40
28	$1.163\ 761\ 4\times10^{-3}$	0.971 145 77	48	$5.562\ 189\ 6\times10^{-5}$	0.973 433 97
29	$7.112\ 053\ 6\times10^{-7}$	0.971 146 49	49	$4.696\ 495\ 0\times10^{-6}$	0.973 438 66
30	$9.830\ 107\ 7\times10^{-4}$	0.972 129 50	50	$2.022\ 646\ 9\times10^{-5}$	0.973 458 89
31	$1.175\ 572\ 3\times10^{-3}$	0.973 305 07	51	$1.510\ 358\ 5\times10^{-6}$	0.973 460 40
32	$2.399\ 591\ 0\times10^{-6}$	0.973 307 47	52	$5.666\ 162\ 4\times10^{-6}$	0.973 466 07
33	$5.971\ 630\ 5\times10^{-6}$	0.973 313 44	53	$1.370\ 555\ 3\times10^{-8}$	0.973 466 08
34	$6.178\ 405\ 3\times10^{-6}$	0.973 319 62	54	$9.392\ 316\ 8\times10^{-6}$	0.973 475 47
35	$4.242\ 487\ 8\times10^{-6}$	0.973 323 86	55	$4.591\ 308\ 4\times10^{-3}$	0.978 066 78
36	$1.993\ 890\ 9\times10^{-6}$	0.973 325 85	56	$3.900\ 383\ 2\times10^{-6}$	0.978 070 68
37	$2.434\ 369\ 4\times10^{-6}$	0.973 328 29	57	$4.682\ 325\ 3\times10^{-6}$	0.978 075 36

可以看出, 此表的最后一列的值和例 6.31 中用直接法得到的相应值是一样的. 当团体保单合同包括了大量个体时, 这种方法特别有用. □

例 6.33 (扩大的群体) 为了说明保单组合规模的影响, 考虑一份包含 1 400 个独立个体的保单组, 这 1 400 个个体是由例 6.30 中的团体寿险保单组合中每个个体变成 100 个个体而得来. 求总损失的精确分布.

解 这里的 n_{ij} 为 100 或 0. 由 (6.42) 式可见, $h(i, k)$ 比原例题中的值大了 100 倍, 总赔付额的分布可以像原例题一样计算, 表 6-23 给出了总赔付额分布函数的一些值 (同前, x 的度量单位是 1 000).

表 6-23 例 6.33 的总概率

x	$F_S(x)$	x	$F_S(x)$	x	$F_S(x)$	x	$F_S(x)$
0	0.007 895 81	200	0.517 933 82	400	0.960 318 65	600	0.999 272 81
8	0.007 895 81	208	0.552 087 36	408	0.965 289 77	608	0.999 392 48
16	0.010 591 83	216	0.575 943 60	416	0.969 998 08	616	0.999 502 70
24	0.019 062 63	224	0.605 836 32	424	0.973 601 99	624	0.999 586 30
32	0.025 613 96	232	0.634 120 23	432	0.976 948 55	632	0.999 655 90
40	0.029 795 97	240	0.666 875 34	440	0.980 318 27	640	0.999 717 84
48	0.037 748 49	248	0.686 324 32	448	0.982 974 84	648	0.999 766 79
56	0.047 703 35	256	0.712 926 41	456	0.985 156 68	656	0.999 808 76
64	0.069 767 56	264	0.739 848 23	464	0.987 281 85	664	0.999 842 05
72	0.076 105 28	272	0.760 610 61	472	0.989 140 69	672	0.999 870 63
80	0.100 134 32	280	0.780 076 67	480	0.990 684 29	680	0.999 894 81
88	0.123 996 72	288	0.800 162 48	488	0.991 948 54	688	0.999 913 71
96	0.138 399 60	296	0.820 737 66	496	0.993 217 71	696	0.999 929 50
104	0.159 456 74	304	0.836 729 20	504	0.994 219 82	704	0.999 942 75
112	0.180 049 88	312	0.851 213 74	512	0.995 046 27	712	0.999 953 64
120	0.221 625 39	320	0.868 491 12	520	0.995 809 92	720	0.999 962 22
128	0.235 697 06	328	0.882 536 29	528	0.996 446 24	728	0.999 969 32
136	0.263 350 63	336	0.893 435 29	536	0.997 014 94	736	0.999 975 45
144	0.301 727 23	344	0.905 305 46	544	0.997 458 96	744	0.999 980 04
152	0.330 416 83	352	0.916 100 01	552	0.997 857 87	752	0.999 983 83
160	0.354 970 38	360	0.925 997 96	560	0.998 219 41	760	0.999 987 07
168	0.386 350 38	368	0.933 408 51	568	0.998 502 04	768	0.999 989 55
176	0.421 778 00	376	0.941 809 39	576	0.998 738 87	776	0.999 991 62
184	0.454 764 09	384	0.949 031 73	584	0.998 950 67	784	0.999 993 26
192	0.478 810 84	392	0.954 829 32	592	0.999 129 71	792	0.999 994 61

由例 6.30 得, 这 1 400 个个体的保单组合总赔付额的均值和方差分别为 $\mu_1 = 205\ 441, \mu_2 = 1.025\ 335\ 6 \times 10^{10}$. 经计算偏度系数为 $\mu_3(\mu_2)^{-3/2} = 0.526\ 734\ 5$, 正好是 14 人团体的相应值的十分之一, 表示分布更加对称了. □

已有大量讨论个体风险模型专题的文献, De Pril[28] 将该方法推广到损失额变

量为多个的情形.

6.11.3 复合 Poisson 近似

由于使用个体风险模型求 n 个风险的保单组合的总赔付额分布的计算十分复杂, 现在通常会尝试使用复合 Poisson 分布来近似总赔付额的分布. 正如 6.5 节所示, 对于复合 Poisson 分布可以用简单的递归过程或用快速傅里叶变换来计算总赔付分布. 总损失的概率生成函数 (6.35) 为

$$P_S(z)=\prod_{j=1}^{n}[1+q_j(z^{b_j}-1)].$$

取对数并对 $\ln[1+q_j(z^{b_j}-1)]$ 进行泰勒级数展开, 得到

$$\ln P_S(z)=\sum_{j=1}^{n}\sum_{k=1}^{\infty}\frac{(-1)^{k+1}}{k}[q_j(z^{b_j}-1)]^k.$$

只保留内部和号的第一项就得到近似公式

$$\ln P_S(z)\doteq\sum_{j=1}^{n}q_j(z^{b_j}-1)=\lambda\sum_{j=1}^{n}\frac{\lambda_j}{\lambda}(z^{b_j}-1), \tag{6.50}$$

其中, $\lambda_j=q_j$ 和 $\lambda=\sum\limits_{j=1}^{n}\lambda_j$, 从而得到

$$P_S(z)\doteq\exp\{\lambda[P_X(z)-1]\},$$

这是复合 Poisson 分布的概率生成函数, 其个体损失分布的概率生成函数为

$$P_X(z)=\frac{1\cdot}{\lambda}\sum_{j=1}^{n}\lambda_j z^{b_j}. \tag{6.51}$$

该分布的概率函数为

$$\Pr(X=x)=\frac{1}{\lambda}\sum_{\{j:b_j=x\}}\lambda_j. \tag{6.52}$$

其分子的求和是对所有保险金额为 b_j 的概率的和.

注意, 索赔赔率分布以及总损失分布的均值都与真实分布相同.

例 6.34 (续例 6.30) *考虑例 6.30 中的团体寿险情形, 推导其复合 Poisson 近似.*

解 使用本节的复合 Poisson 近似法, 参数 $\lambda=\sum q_j=0.048\ 13$, 进而得到表 6-24 所示的分布函数.

将这些值与例 6.31 的结果比较, 可以发现最大的误差是 0.000 270 8, 发生在 $x=0$ 点. □

表 6-24 例 6.34 的总分布

x	$F_S(x)$	x	$F_S(x)$	x	$F_S(x)$	x	$F_S(x)$
0	0.953 009 9	20	0.961 834 8	40	0.973 577 1	60	0.999 097 4
1	0.953 009 9	21	0.961 834 8	41	0.973 585 0	61	0.999 098 6
2	0.953 009 9	22	0.961 834 8	42	0.973 607 2	62	0.999 099 4
3	0.953 009 9	23	0.961 834 8	43	0.973 613 3	63	0.999 099 5
4	0.953 009 9	24	0.966 447 3	44	0.973 634 6	64	0.999 099 5
5	0.953 009 9	25	0.966 447 3	45	0.973 639 3	65	0.999 099 6
6	0.953 009 9	26	0.970 202 2	46	0.973 651 3	66	0.999 099 7
7	0.953 009 9	27	0.970 202 2	47	0.973 654 1	67	0.999 099 7
8	0.953 009 9	28	0.971 365 0	48	0.973 670 8	68	0.999 099 8
9	0.953 009 9	29	0.971 365 7	49	0.973 675 5	69	0.999 102 2
10	0.953 009 9	30	0.972 349 0	50	0.973 695 6	70	0.999 109 1
11	0.953 009 9	31	0.973 523 5	51	0.973 697 1	71	0.999 115 6
12	0.953 009 9	32	0.973 526 8	52	0.973 710 1	72	0.999 117 9
13	0.953 009 9	33	0.973 532 8	53	0.973 710 2	73	0.999 134 1
14	0.953 486 4	34	0.973 539 1	54	0.973 719 5	74	0.999 147 0
15	0.954 906 4	35	0.973 543 3	55	0.978 290 1	75	0.999 183 9
16	0.956 259 7	36	0.973 551 2	56	0.978 294 7	76	0.999 213 5
17	0.956 736 2	37	0.973 553 6	57	0.978 299 4	77	0.999 223 9
18	0.960 100 3	38	0.973 560 4	58	0.978 300 6	78	0.999 297 3
19	0.960 614 9	39	0.973 567 9	59	0.978 302 1	79	0.999 330 7

也可以使用其他类似的近似方法, 比较常见的是令 (6.50) 式中的 λ_j 为

$$\lambda_j = -\ln(1-q_j), \quad j=1,2\cdots,n. \tag{6.53}$$

这使得损失不发生的概率 $1-q_j$ 与 Poisson 分布损失不发生的概率 $\mathrm{e}^{-\lambda_j}$ 相匹配, 用 Poisson 分布有效地替代了群体中的每个个体分布. 这种近似适用于团体寿险合同, 当其中的某个个体因死亡被 "替换" 时, 团体的 Poisson 死亡强度仍保持不变. 显然, 损失次数的期望值大于 $\sum\limits_{j=1}^{n} q_j$. Kornya[79] 中提出了另一种方法, 在 (6.50) 式中令 $\lambda_j = q_j/(1-q_j)$, 这导致损失次数的期望值超过了使用 (6.53) 式时的期望值. (见习题 6.61).

损失额变量不唯一的情形

本节开始已经提到, 损失额可能存在不止一个变量的情况. 同样令 B_j 为给定损失发生的条件下度量损失额的随机变量, 令 $X_j = I_j B_j$, 则

$$P_{X_j}(z) = [1-q_j+q_j P_{B_j}(z)].$$

对应于 (6.35) 式的概率生成函数为

$$P_S(z)=\prod_{j=1}^{n}[1-q_j+q_jP_{B_j}(z)].$$

尽管可以将精确计算方法推广应用到这里, 但是十分繁琐. 相反, 基于矩匹配的复合 Poisson 近似 (6.50) 式仅仅需要用 $P_{B_j}(z)$ 替代 z^{b_j}. 从而, 损失分布的概率生成函数 (6.51) 式变成

$$P_X(z)=\frac{1}{\lambda}\sum_{j=1}^{n}\lambda_jP_{B_j}(z),$$

所以

$$f_X(x)=\frac{1}{\lambda}\sum_{j=1}^{n}\lambda_jf_{B_j}(x), \tag{6.54}$$

为 n 个个体损失程度密度的加权平均. 将其推广到连续损失分布的情形, 对所有的 x 值 (6.54) 式同样是成立的.

例 6.35 (续例 6.29)　*用以上 3 种方法求复合 Poisson 近似分布, 计算每种近似分布的均值和方差, 并与精确值作比较.*

解　使用均值匹配法, 得到 $\lambda=50(0.01)+25(0.01)=0.75$, 损失分布为

$$\begin{aligned}
f_X(50\ 000)&=\frac{50(0.01)(0.7)}{0.75}=0.466\ 7,\\
f_X(75\ 000)&=\frac{25(0.01)(0.7)}{0.75}=0.233\ 3,\\
f_X(100\ 000)&=\frac{50(0.01)(0.3)}{0.75}=0.200\ 0,\\
f_X(150\ 000)&=\frac{25(0.01)(0.3)}{0.75}=0.100\ 0.
\end{aligned}$$

均值为 $\lambda\mathrm{E}(X)=0.75(75\ 833.33)=56\ 875$, 和真实值一致, 方差为 $\lambda\mathrm{E}(X^2)=0.75(6\ 729\ 166\ 677)=5\ 046\ 875\ 000$, 超出了真实值.

对于保持损失不发生的概率值不变的方法,$\lambda=-75\ln(0.99)=0.753\ 775$. 在这种方法下的损失分布也是与前面完全一致的 (这是由于所有个体都有相同的 q_j 值). 因此均值为 57 161, 方差为 5 072 278 876, 两者都超出了前一种方法的近似值.

使用 Kornya 方法, $\lambda=75(0.01)/0.99=0.757\ 576$, 损失分布也没有改变. 均值为 57 449, 方差为 5 097 853 535, 两者都是三种方法中最大的.　□

习题

6.60　推导 (6.36) 式和 (6.37) 式.

6.61 证明：给定 $\lambda_j = q_j$ 和 (6.52) 式时，复合 Poisson 模型产生的均值与精确分布相同，但方差会偏大. 再证明：使用 $\lambda_j = -\ln(1-q_j)$ 时，必会产生比复合 Poisson 模型更大的均值和方差. 最后证明：使用 $\lambda_j = q_j/(1-q_j)$ 时，产生的均值和方差比前两者都大.

6.62 某团体保单中每个成员的赔案相互独立，索赔额分布的统计数据由表 6-25 给出.

设团体的未来赔付额为 S，保费为 S 的均值加上其标准差的两倍. 已知，若团体的 m 位成员性别未知时，假设男性成员数服从参数为 m 且 $q = 0.4$ 的二项分布. 令 A 为由 100 位性别未知成员组成的团体的保费，B 为 40 位男性和 60 位女性组成的团体的保费. 求 A/B.

表 6-25 习题 6.62 的数据

	均值	方差
男性	2	4
女性	4	10

6.63 保险公司假设团体寿险合同覆盖人群的索赔概率如表 6-26 所示.

表 6-26 习题 6.63 的数据

类别	索赔概率
吸烟者	0.02
非吸烟者	0.01

现有个体相互独立的团体，其中每个个体的投保额为 1 000 元. 保险公司假设 20%的个体是吸烟者. 基于此假设，将保费定为期望赔付额的 110%. 如果 30%的个体是吸烟者，则索赔额超过保费的概率小于 0.2. 使用正态近似，计算该团体包含的最少个体数.

6.64 基于赔案相互独立的个体风险模型，表 6-27 给出了某寿险保单组合总赔付额的累积分布函数.

表 6-27 习题 6.64 的分布

x	$F_S(x)$
0	0.40
100	0.58
200	0.64
300	0.69
400	0.70
500	0.78
600	0.96
700	1.00

现将一份原保险金额为 100, 索赔概率 0.2 的保单的保险金额增加到 200, 求作此修改后保单组合的总赔付额超过 500 的概率.

6.65 某团体寿险合同的个体是独立的, 且该团体被分为 3 个年龄段, 如表 6-28 所示. 已知合同的定价使得总赔付额超过保费的概率为 0.05. 使用正态近似, 求保险人收取的保费.

表 6-28 习题 6.65 的数据

年龄段	年龄段的人数	个体索赔概率	指数分布的索赔额的均值
18~35	400	0.03	5
36~50	300	0.07	3
51~65	200	0.10	2

6.66 某健康保障计划中每个成员每年的索赔额分布概率模型如表 6-29 所示. 假设每位成员接受服务的频率和服务费用是独立的, 并且各成员相互独立. 使用正态近似, 求该保障计划至少包含多少成员, 才能使真实费用超过费用期望值的 115%的概率小于 0.10.

表 6-29 习题 6.66 的数据

服务	索赔概率	给定索赔发生的条件下年费用的分布	
		均值	方差
门诊	0.7	160	4 900
手术	0.2	600	20 000
其他服务	0.5	240	8 100

6.67 某保险公司现有如表 6-30 所示的风险独立的保单组合. 保险公司设定 α 和 k 值使得总赔付额的期望值为 100 000, 且方差最小. 求 α.

表 6-30 习题 6.67 的数据

类别	索赔概率	受益	风险体个数
标准	0.2	k	3 500
非标准	0.6	αk	2 000

6.68 保险公司拥有一组一年期风险独立的寿险保单组合, 如表 6-31 所示. 该保险公司的精算师使用复合 Poisson 模型来近似计算个体风险模型中的索赔额分布, 其中 Poisson 模型的期望赔案数与个体模型相同. 求 x 的最大值使得复合 Poisson 近似分布的方差小于 4 500.

表 6-31 习题 6.68 的数据

类别	类别包含的个体数	受益额	索赔概率
1	500	x	0.01
2	500	$2x$	0.02

6.69 保险公司出售了一组共 2 300 份独立个体的一年期寿险保单, 如表 6-32 所示.

表 6-32 习题 6.69 的数据

类别	受益额	死亡概率	保单数目
1	100 000	0.10	500
2	200 000	0.02	500
3	300 000	0.02	500
4	200 000	0.10	300
5	200 000	0.10	500

保险公司对每个个体索赔额超过 100 000 的部分进行了再保险. 再保险公司希望收到的保费可以保证亏损的可能性只有 5%. 用下列每种方法计算合理的保费.

(a) 对总索赔额分布使用正态近似.
(b) 使用对数正态近似.
(c) 使用 gamma 近似.
(d) 使用均值匹配的复合 Poisson 近似.
(e) 进行精确计算 (使用 De Pril 提出的方法或者其他方法). 这需要编程计算.

第 7 章 离散时间破产模型

7.1 引 言

虽然很难对保险合同的组合风险进行评估, 但这项工作对保险业务的可行性却是非常重要的. 显然, 某特定时期总赔付额的分布是评估过程中一个重要的考虑变量, 这类变量的分布是前几章所阐述的主要内容.

本章采用多期的方法来跟踪研究每个保单、业务组合或者整个公司的财富随时间的变化趋势. 最常用的方法是**破产理论**. 该理论最关心的量是盈余, 当盈余变为负值时将发生破产. 为了跟踪盈余的变化情况, 模型中不能只考虑索赔的支付, 还需要包括保费、投资收入和各种费用以及任何对现金流产生影响的项目.

为了保证数学处理的简单, 本章以及第 8 章描述的模型都是相当简化和理想化的情形. 因此, 不应该将分析结果视为对绝对现实的表现. 但它的确能够对公司业务组合的风险提供重要的附加信息, 这些信息对于公司的长期财务规划以及维持保险公司的偿付能力是有意义的.

本章由两部分组成. 第一部分 (7.2 节) 介绍过程模型. 这一部分将给出一些基本概念以及破产理论主要术语的定义. 第二部分 (7.3 节) 分析离散时间模型. 对这个模型的分析要用到前面章节的一些工具, 而关于连续时间模型的分析会在第 8 章讨论, 那时, 需要用到随机过程的知识. 我们将主要从分析以下 2 个过程入手: 复合 Poisson 过程和布朗运动. 在精算学领域, 复合 Poisson 过程是进行破产分析的标准模型, 而在现代金融理论中, 布朗运动同样具有相当大的作用, 另外, 布朗运动还可以用作复合 Poisson 过程的一种近似.

7.2 保险过程模型

7.2.1 过程

本章与前几章最主要的不同的是, 我们现在需要考察业务组合随时间的演变过程. 为此我们定义 2 种过程. 注意到, 那些包含随机事件的过程通常称作**随机过程**, 但我们不准备使用修饰词 "随机的", 而是相信通过上下文可以清楚地表明过程是否为随机的.

定义 7.1 连续时间过程,记作 $\{X_t; t \geqslant 0\}$. 如果存在一些随机元素,使得对所有的

$t_1, \cdots, t_n$ 及任一个 n 都可以具体给出 $(X_{t_1}, \cdots, X_{t_n})$ 的联合分布.

一般说来, 只有 X_t(对任意的 t) 的具体分布还不足以描述过程的性质. 许多过程在不同时间点的观测值之间还存在相关关系.

例 7.2 设 $\{S_t; t \geqslant 0\}$ 为时间 0 到时间 t 的总损失量, 指出如何用第 6 章中的聚合风险模型描述这个过程.

解 设 $t_1 < \cdots < t_n$, 对于 $(S_{t_1}, \cdots, S_{t_n})$ 的联合分布, 取 $W_j = S_{t_j} - S_{t_{j-1}}$, 且 $S_{t_0} = S_0 = 0$. 设 W_j 可以用聚合风险模型表示而且相互独立. 当频率分布的均值与时间长度 $t_j - t_{j-1}$ 成比例时, 个体损失分布将会完全相同. 图 7-1 为该过程的一次实现, 称其为一条**样本轨道**. □

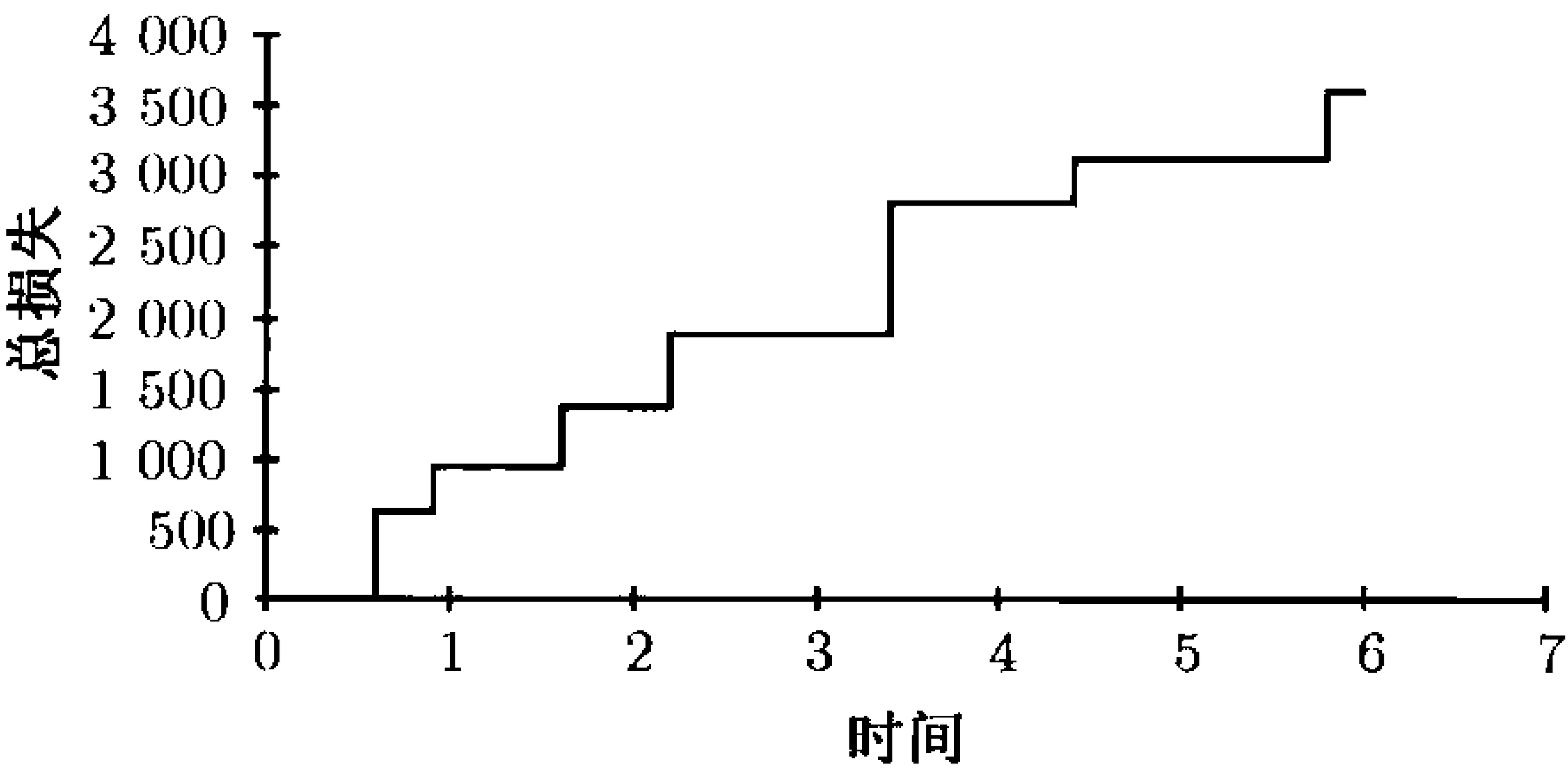


图 7-1 连续的总损失过程, S_t

通常情况下, 如果某个过程随着时间的推移没有很大的变化, 则这样的过程更容易描述. 以下是两种具体的定义方法.

定义 7.3 称过程满足独立增量: 如果对任意 $s < t \leqslant v < u$, 随机变量 $X_t - X_s$ 与 $X_u - X_v$ 独立.

这个性质指出了在任意不重叠的时期, 过程的变动是独立的.

定义 7.4 称过程满足平稳增量: 若 $X_t - X_s$ 的分布仅仅由 t 与 s 的差值 $t-s$ 决定.

这个性质暗示过程的变动不由所在的时刻决定. 换句话说, 仅仅观察过程的增量本身你无法断定它的发生时刻.

大多数商业机构都无法连续的监控其经营状况. 取而代之, 一般采取定期检查的方法. 为此, 我们引出其他的各种过程.

定义 7.5 离散时间过程, 记作 $\{X_t; t = 0, 1, 2, \cdots\}$. 如果存在一些随机元素, 使得对任意整数 t_i 及 n 都可以得到 $(X_{t_1}, \cdots, X_{t_n})$ 的联合分布.

离散时间过程可以看作是一个连续时间过程仅仅在整数时刻记下了 X_t 的值. 本章所有的离散时间过程都是在观测时期末进行计量的, 例如每月、每季或每年.

例 7.6 (续例 7.2) 将例7.2的过程转换成一个含有平稳独立增量的离散时间过程.

解 令 $X_1, X_2, \cdots$ 为每个时期的总损失量, 其中$\{X_j\}$是独立同分布的, 每一个 X_j

为复合分布, 记总损失过程为 $S_t = X_1 + \cdots + X_t$. 此过程具有平稳增量性, 因为 $S_t - S_s = X_{s+1} + \cdots + X_t$, 而且它的分布仅仅取决于 X_j 的个数即 $t-s$. 独立增量性则直接由$\{X_j\}$的独立性得到. 图 7-2 为图 7-1 的离散时间模型. □

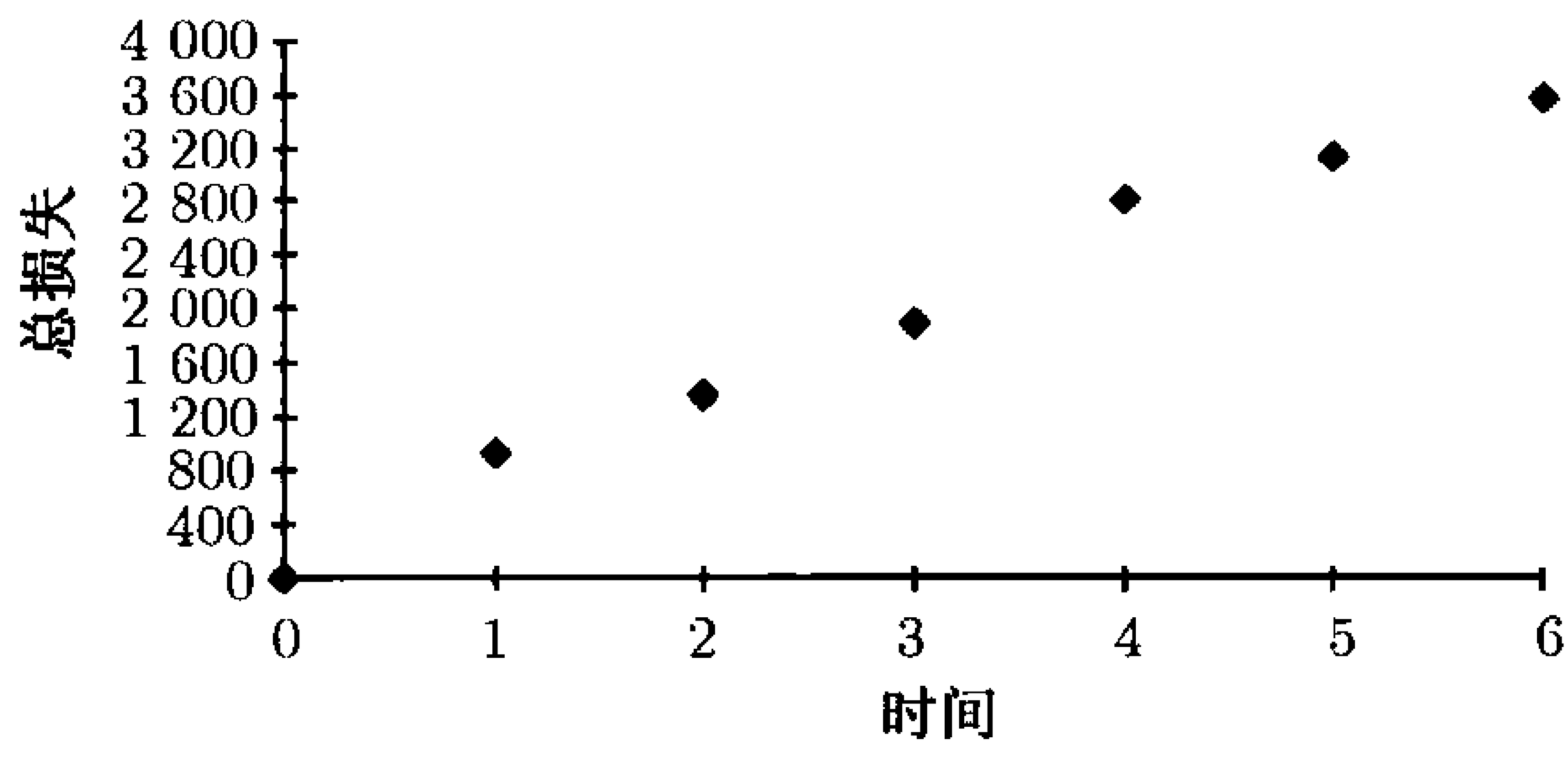


图 7-2 离散的总损失过程, S_t

7.2.2 保险模型

前面章节讨论的一些例子已经阐释了保险过程所采用的大部分模型. 我们关心的是**盈余过程**$\{U_t; t \geqslant 0\}$(或离散情形 $\{U_t; t = 0, 1, \cdots\}$), 它衡量了业务组合到 t 时刻为止的盈余. **设初始盈余**$u = U_0$. 这里是从会计的角度来考虑盈余, 表示若业务组合在该时刻终止, 可能出现的资金盈余. 从持续经营的角度看, 出现正的盈余将表示公司有一定的抗风险能力. 时刻 t 的盈余定义为

$$U_t = U_0 + P_t - S_t,$$

其中 $\{P_t; t \geqslant 0\}$ 为保费过程, 表示时刻 t 之前收取的所有保费 (净保费), $\{S_t; t \geqslant 0\}$ 为损失过程, 表示时刻 t 之前的所有赔付. 因此, 有以下的说明:

(1) P_t 可以表示签单保费也可以表示已赚保费, 视情况而定;

(2) S_t 可以表示已记录的赔付或表示已发生的损失, 同样视情况而定;

(3) 当 $u < t$ 时, P_t 可能会依赖于 S_u. 例如, 基于过去良好的损失经验可能会适当的减少当前的保费以此作为奖励.

虽然不一定必要, 但是可以将 S_t 的频率部分与个体损失量部分分开. 记$\{N_t; t \geqslant 0\}$ 为赔付次数过程, 记录了时刻 t 之前的赔付次数. 记 $S_t = X_1 + \cdots + X_{N_t}$, 不一定要求 $\{X_1, X_2, \cdots\}$ 满足独立同分布条件. 然而, 如果它们是独立同分布的, 且与 N_t 独立 (对所有的 t), 则 S_t 将满足复合分布.

下面考虑两种特殊情形的盈余过程. 本章只讨论这两种过程.

离散时间模型

定义盈余过程在第 t 年的增量为

$$W_t = P_t - P_{t-1} - S_t + S_{t-1}, \quad t = 1, 2, \cdots,$$

然后进一步得到

$$U_t = U_{t-1} + W_t, \quad t = 1, 2, \cdots.$$

如此说来, 只要随机变量 W_t 相对其他任一个 W_t 是独立的, 或仅仅取决于 U_{t-1} 的值, 那么求解 $\{U_t; t = 0, 1, 2, \cdots\}$ 的分布会相对容易. W_t 对 U_{t-1} 的依赖性使我们可以通过当年年末盈余的计算来分配红利 (因为 W_t 取决于 P_t).

7.3 节将介绍两种确定 U_t 分布的方法. 计算量会很大, 但只要有足够的时间和资源, 答案是容易获得的.

连续时间模型

在大多数情形下分析连续时间模型是极其困难的. 这是因为必须在每一个时间点上建立联合分布, 而不是仅仅建立在可数个时间点上. 广泛讨论的一个模型是复合 Poisson 索赔过程, 其中保费以固定的非随机的比例连续收取

$$P_t = (1+\theta)\mathrm{E}(S_1)t,$$

总的损失过程为

$$S_t = X_1 + \cdots + X_{N_t},$$

其中 $\{N_t; t \geqslant 0\}$ 为 Poisson 过程. 这个过程将在 8.1.1 节中详细介绍. 现在只要知道, 该模型任何一段时期中的损失次数满足 Poisson 分布, 并且它的均值与时间段的长度成比例.

因为这类模型的分析将更为困难, 整个第 8 章将完全用来讨论这个模型的推导和分析. 现在我们开始定义一个非常关心的变量, 这是一个能够衡量业务组合成功可能性的量.

7.2.3 破产

建立过程模型的主要目的是确定一个业务组合生存的可能性. 可以用 4 种方法定义这里的生存概率.

定义 7.7 *连续时间永久生存概率, 定义为*

$$\phi(u) = \Pr(U_t \geqslant 0, \text{ 对所有 } t \geqslant 0 | U_0 = u).$$

为此, 我们需要持续不断地检查公司的盈余, 并要求业务组合永远具有偿债能力. 事实上, 持续不断的检查以及要求业务组合永远存在都是不切实际的. 在实践中, 更可能采取的方法是定期检查盈余. 人们自然希望经营能够永远持续下去, 然而要求模型能够预测无限远的未来似乎过于苛刻. 下面介绍一个更有用的量.

定义 7.8 离散时间有限生存概率, 定义为

$$\widetilde{\phi}(u,\tau)=\Pr(U_t\geqslant 0,\ \text{对所有}\ t=0,1\cdots,\tau|U_0=u).$$

它度量了业务组合生存到 τ 时刻 (通常用年表示) 的概率, 并且只需要在每个时期末检查是否破产. 此外, 还有两种描述中间情形的量.

定义 7.9 连续时间有限生存概率, 定义为

$$\phi(u,\tau)=\Pr(U_t\geqslant 0,\ \text{对所有}\ 0\leqslant t\leqslant\tau|U_0=u),$$

离散时间永久生存概率, 定义为

$$\widetilde{\phi}(u)=\Pr(U_t\geqslant 0,\ \text{对所有}\ t=0,1,\cdots,|U_0=u).$$

以下不等式的成立是显然的:

$$\widetilde{\phi}(u,\tau)\geqslant\widetilde{\phi}(u)\geqslant\phi(u),$$

及

$$\widetilde{\phi}(u,\tau)\geqslant\phi(u,\tau)\geqslant\phi(u).$$

极限情境也应该是显而易见相等的, 如

$$\lim_{\tau\to\infty}\phi(u,\tau)=\phi(u),$$

$$\lim_{\tau\to\infty}\widetilde{\phi}(u,\tau)=\widetilde{\phi}(u).$$

在许多情形下, 收敛是迅速的. 这意味着对有限或永久的选择应该同时取决于计算的简便性以及模型的适用性. 我们发现, 在 Poisson 过程情境中永久概率相对容易获得. 而对于另一些情形, 有限时间的计算应该更容易一些.

尽管没有用专门的符号去表示, 但是还有另外的一对极限关系成立. 当盈余检查的频率逐渐增大 (这里指每年的次数), 离散时间的生存概率会收敛于其所对应的连续时间情形.

既然本小节以破产为主题, 这里就以破产概率的定义作为结束.

定义 7.10 连续时间无限期破产概率, 定义为

$$\psi(u)=1-\phi(u).$$

另外 3 个破产概率的定义也可以用类似的方法得到.

7.3 离散时间有限破产概率

7.3.1 离散时间过程

令 P_t 表示第 t 期收到的保费, S_t 表示第 t 期的赔付. 也可以考虑更一般的情形, 令 C_t 为保险费和赔付以外的任意现金流, 最主要的这类现金流是期初可支配盈余的当期投资所得. 因此, 第 t 期末的盈余为

$$U_t = u + \sum_{j=1}^{t}(P_j + C_j - S_j) = U_{t-1} + P_t + C_t - S_t.$$

最后还有一个假设是: 给定 U_{t-1} 时随机变量 $W_t = P_t + C_t - S_t$ 仅仅取决于 U_{t-1}, 而不依赖于任何先前的信息. 称这样的 $\{U_t; t = 1, 2, \cdots\}$ 为**Markov 过程**.

为了计算破产概率, 我们考虑如下定义的第二过程. 首先, 定义

$$\begin{aligned} W_t^* &= \begin{cases} 0, & U_{t-1}^* < 0, \\ W_t, & U_{t-1}^* \geqslant 0, \end{cases} \\ U_t^* &= U_{t-1}^* + W_t^*, \end{aligned} \tag{7.1}$$

其中的新过程以 $U_0^* = u$ 为初值, 这种情形下, 有限生存概率为

$$\widetilde{\phi}(u, \tau) = \Pr(U_\tau^* \geqslant 0).$$

因此, 只需要检查 τ 时刻的 U_t^*, 因为一旦破产, 该过程就再也不会为非负值. 接下来的例子说明了这个特性, 也是对 7.3.2 节将详细介绍的方法的初步说明.

例 7.11 考虑这样一个初始盈余为 2 的过程, 固定的年保费为 3, 损失为 0 或 6 的概率分别为 0.6 和 0.4. 除此以外没有其他现金流, 试计算 $\widetilde{\phi}(2,2)$.

解 这里的 U_1 仅有 2 种可能值: 5 和 −1, 概率分别为 0.6 和 0.4. 每年的 W_t 取值为 3 或 −3, 概率分别为 0.6 和 0.4. 在第 2 年底过程有四种可能的结果, 它们都列在表 7-1 中. 则 $\widetilde{\phi}(2,2)$=0.36+0.24=0.6. 为了考虑 U_2 还需要分析情形 3 和情形 4, 分别得到 2 和 −4. 但这里的过程不考虑从破产中恢复过来的情况, 所以必须将情形 3 一直设为负数. □

表 7-1 例 7.11 的计算结果

情形	U_1	W_2	W_2^*	U_2^*	概率
1	5	3	3	8	0.36
2	5	−3	−3	2	0.24
3	−1	3	0	−1	0.24
4	−1	−3	0	−1	0.16

7.3.2　计算破产概率

一般有 3 种方法计算破产概率. 一种永远可行的是随机模拟, 正如可以对总损失分布进行模拟一样, 也可以对盈余过程进行同样的模拟. 对于极其复杂的模型(例如, 医疗险的赔付模型, 不仅包含住院费、处方药和出诊费用等, 还要考虑通货膨胀、利率和使用率的变化), 这种方法可能是唯一可行的. 对于一些相对简单的模型, 可以使用其他 2 种方法并且效果不错. 第一种是直接计算的方法, 只有很少的限制条件, 第二种是反演计算, 需要一些限制条件.

卷积计算

在实际应用中这种方法通常要求所有随机变量都是离散的, 而且支集为有限集合. 如果不是这样的, 应该首先进行离散近似. 然后利用公式 (7.1) 递推计算. 为了符号的方便, 假定我们已经得到了 U_{t-1}^* 的离散概率函数. 则破产概率为 $\widetilde{\psi}(u,t-1) = \Pr(U_{t-1}^* < 0)$. 非负盈余的分布为 $f_j = \Pr(U_{t-1}^* = u_j)$, $j = 1,2,\cdots,n$, 其中对任意的 j, $u_j \geqslant 0$, 并且 u_n 为 U_{t-1}^* 的最大值. 已假设对于给定的 U_{t-1}^* 的每一个正值, W_t 的分布也是已知的.

令 $g_{j,k} = \Pr(W_t = w_{j,k}|U_{t-1}^* = u_j)$, 也不排除 W_t 依赖于 u_j 的可能. 然后利用卷积方法计算 U_t^* 的概率. 首先,

$$\begin{aligned}
\widetilde{\psi}(u,t) &= \widetilde{\psi}(u,t-1) + \Pr(U_{t-1}^* \geqslant 0 \ \text{且}\ U_{t-1}^* + W_t < 0) \\
&= \widetilde{\psi}(u,t-1) + \sum_{j=1}^{n} \Pr(U_{t-1}^* + W_t < 0|U_{t-1}^* = u_j)\Pr(U_{t-1}^* = u_j) \\
&= \widetilde{\psi}(u,t-1) + \sum_{j=1}^{n} \Pr(u_j + W_t < 0|U_{t-1}^* = u_j) f_j \\
&= \widetilde{\psi}(u,t-1) + \sum_{j=1}^{n} \sum_{w_{j,k} < -u_j} g_{j,k} f_j.
\end{aligned}$$

然后, 有

$$\begin{aligned}
\Pr(U_t^* = x) &= \Pr(U_{t-1}^* \geqslant 0 \ \text{且}\ U_{t-1}^* + W_t = x) \\
&= \sum_{j=1}^{n} \Pr(U_{t-1}^* \geqslant 0 \ \text{且}\ U_{t-1}^* + W_t = x|U_{t-1}^* = u_j) \times \Pr(U_{t-1}^* = u_j) \\
&= \sum_{j=1}^{n} \Pr(u_j + W_t = x|U_{t-1}^* = u_j) f_j \\
&= \sum_{j=1}^{n} \sum_{w_{j,k} + u_j = x} g_{j,k} f_j.
\end{aligned}$$

尽管这些公式看上去很繁琐，但还是很容易编程实现的，考虑下面的例子.

例 7.12 (题中所有的数量金额都是以适当的货币单位表示的) 假设年度总损失的取值为 0,2,4,6 的概率分别为 0.4,0.3,0.2,0.1. 进一步假设初始盈余为 2 元，每年初收到的保险费为 2.5 元. 每年的利息都是年初可支配盈余的 10%，索赔在年末赔付. 另外，如果当年没有发生损失，将返还 0.5 元的保险费. 计算前 2 年年末的生存概率.

解 首先注意保费返还不能抵免下一年的保费. 这样，在每年初不仅需要盈余量，还要知道当年是否有保费返还.

在 0 时刻，有 $\widetilde{\psi}(2,0)=0$ 和 $f_1=\Pr(U_0^*=2)=1$. 下面的表 7-2 给出了 $w_{1,k}$ 的可能取值和概率 $g_{1,k}$.

表 7-2 例 7.12 的 w 和 g

k	$w_{1,k}$	$g_{1,k}$
1	2.45	0.4
2	0.95	0.3
3	−1.05	0.2
4	−3.05	0.1

例如，$w_{1,1}$ 将包含保险费 2.5、利息 0.45(收取保险费后的盈余)、赔付 0 和保费返还 0.5. 为了求 $\widetilde{\psi}(2,1)$ 的值，注意到 $w_{1,k}$ 的值当 $-u_1=-2$ 时只有 $w_{1,4}$，即 $\widetilde{\psi}(2,1)=0.1$. 也容易看出，取正概率的情形就是那些使 $2+w_{1,k}>0$ 成立的 x 值. 下面的表 7-3 给出了 U_1^* 的分布.

表 7-3 例 7.12 的 U_1^*

j	u_j	f_j
1	0.95	0.2
2	2.95	0.3
3	4.45	0.4

剩余的概率为 $\Pr(U_1^*=-1.05)=0.1$.

为了使第 2 年的情景可视化，可采用一个二维表给出所有关于 u_j 和 $w_{j,k}$ 的可能组合. 表 7-4 中的数据为 $u_j+w_{j,k}$ 和 $g_{j,k}$，因为我们只关心这些求和量，所以这里只给出了它们的和.

表中每个单元格的联合概率是其所在行的 f_j 与单元格中列出的概率值的乘积. 除 $\widetilde{\psi}(2,1)$ 外，还要考虑表 7-4 中所有含负数的单元格的概率，即

$$\widetilde{\psi}(2,2)=0.1+0.2(0.2)+0.2(0.1)+0.3(0.1)=0.19.$$

这里不能重复 $w_{i,k}$ 的值，所以最好给这些值排序，并注意到它们就是第 3 年年初的

新 u_j. 表 7-5 列出了这些值.

表 7-4 例 7.12 的 $u+w$ 和 g

j	u_j	f_j	k			
			1	2	3	4
1	0.95	0.2	3.295,0.4	1.795, 0.3	−0.205, 0.2	−2.205, 0.1
2	2.95	0.3	5.495, 0.4	3.995, 0.3	1.995, 0.2	−0.005, 0.1
3	4.45	0.4	7.145, 0.4	5.645, 0.3	3.645, 0.2	1.645, 0.1

表 7-5 例 7.12 的 u

j	u_j	f_j
1	1.645	0.04
2	1.795	0.06
3	1.995	0.06
4	3.295	0.08
5	3.645	0.08
6	3.995	0.09
7	5.495	0.12
8	5.645	0.12
9	7.145	0.16

总概率为 0.81, 为 $\widetilde{\psi}(2,2)$ 的余数 (两个和为 1 的数互为余数). 由更早的定义, 剩余的概率 0.19 是与 $U_2^* < 0$ 相联系的.

易于发现, u 的取值个数以及这些值的小数后位数将迅速地增加. 在某些情形, 取整运算是个好方法. 简单的方法是, 要求每个时期的 u 值是某个 h 的倍数, 这个跨度可能要随着时间推移而增大. 如果在某个非 h 倍数的位置存在正概率, 可以将它分配到 2 个最近的 h 倍数的位置, 并保持均值不变. (分配到更多的位置可以保持更高阶的矩). □

例 7.13 (续例 7.12) *用跨度 $h=2$, 调整第 2 年末的盈余的概率分布.*

解 在 1.645 点的概率为 0.04, 应该将其分配到 0 和 2 上. 为了保持均值不变, 可以将 $0.355*0.04/2=0.007\,1$ 分配在 0 点上, 剩余的 0.032 9 分配在点 2 上. 则期望值为 $0.007\,1*0+0.032\,9*2=0.065\,8$, 与原值 $0.04*1.645=0.065\,8$ 相同. 计算中的 0.355 为 1.645 与 2 的距离, 分母为定义的跨度. 在区间的左端点上放置了这个概率. 得到的近似分布最后在表 7-6 中给出. □

反演计算

反演方法的一个优势是将卷积计算简化为乘法计算, 只要随机变量是相互独立的即可. 这里意味着 W_t 与 U_{t-1} 独立. 我们用另一种方法来追踪破产的轨迹 (前面的方法是冻结破产时刻的 U_t^*). 这个想法也可以应用在直接的卷积计算中. 令 U_t^{**}

为 $U_t \geqslant 0$ 条件下的 U_t 值. 在每个期末, 所有与破产相关的概率将重新分配在非负盈余的结果之中. 用如下方法逐年进行分析:

(1) 计算 U_{t-1}^{**} 的特征函数 $\varphi_{1,t}(z) = \mathrm{E}(\mathrm{e}^{\mathrm{i}zU_{t-1}^{**}})$;

(2) 计算 W_t 的特征函数 $\varphi_{2,t}(z) = \mathrm{E}(\mathrm{e}^{\mathrm{i}zW_t})$;

(3) 然后有 $U_{t-1}^{**} + W_t$ 的特征函数为 $\varphi_{3,t}(z) = \varphi_{1,t}(z)\varphi_{2,t}(z)$;

(4) 用反演方法确定 $U_{t-1}^{**} + W_t$ 的概率函数 $f_t(u)$;

(5) 令 $r_t = \Pr(U_{t-1}^{**} + W_t < 0)$. 这是已知生存到时刻 $t-1$ 的条件下, 业务组合在时刻 t 破产的概率;

(6) 则 $f_t^{**}(u) = f_t(u)/(1-r_t), u \geqslant 0$ 为 U_t^{**} 的概率函数;

(7) 时刻 t 的破产概率为: $\widetilde{\psi}(u,t) = \widetilde{\psi}(u,t-1) + r_t[1-\widetilde{\psi}(u,t-1)]$.

上述过程的第一步由 $U_1 = u + W_1$ 直接得到 U_1 的概率函数, 只要将 W_1 的概率函数的对应项平移 u 即可.

表 7-6 例 7.13 的概率

j	u_j	f_j
1	0	0.013 4
2	2	0.189 225
3	4	0.258 975
4	6	0.256 8
5	8	0.091 6

例 7.14 已知年总损失为 0, 2, 4, 6 的概率分别为 0.4, 0.3, 0.2, 0.1. 每年年初收到的保险费为 2.5, 初始盈余为 2. 用快速 Fourier 变换计算在 2 年内破产的概率.

解 W_t 在各年的概率函数相同, 列在表 7-7 中. 由初始盈余为 2, 可以很容易得到 U_1 的分布. 由表 7-8 给出. 这就直接得到 $\widetilde{\psi}(2,1) = 0.1, U_1^{**}$ 的分布在表 7-9 中给出.

表 7-7 例 7.14 W_t 的概率函数

w	$\Pr(W=w)$
−3.5	0.1
−1.5	0.2
0.5	0.3
2.5	0.4

表 7-8 例 7.14 U_1 的概率函数

u	$\Pr(U_1=u)$
−1.5	0.1
0.5	0.2
2.5	0.3
4.5	0.4

表 7-9 例 7.14 U_1^{**} 的概率函数

u	$\Pr(U_1^{**}=u)$
0.5	2/9
2.5	3/9
4.5	4/9

为了使 FFT(快速 Fourier 变换) 操作简单, 最好每一个量都是正值. 为了做到这一点只需将每个值加上 3.5 即可. 变化后的分布列在表 7-10 的第 2 列和第 3 列. 如果预见到变换 $U_1^{**}+W_2$ 从 0 到 14 每间隔 2 取一次值, 因此需要 8 个值. 这已经是 2 的幂形式, 不用再补零. 表 7-10 的第 4 列和第 5 列是输入 (U_1^{**}、W_2) 后的 FFT 变换结果. 后面紧接着的两列分别是: 两个特征函数的乘积和这个乘积的逆. 最后一列就是要求的概率函数. 当然, 这个例子用卷积方法计算也是比较简单的, 但这里的计算说明 FFT 确实可以达到预期的目的.

我们必须注意表 7-10 中最后一列的概率是平移了 7 之后的结果, 实际分布由表 7-11 给出. 注意到有 9/90=1/10 取负值的概率, 所以有 $\widetilde{\psi}(2,2)=0.1+0.9*0.1=0.19$. 条件分布 U_2^{**} 也列在表 7-11 中. □

表 7-10 例 7.14 第二年的破产概率

u	$f_1^{**}(u)$	$f_W(u)$	$\varphi_{1,2}/8$	$\varphi_{2,2}/8$
0	0	1/10	0.125	0.125
2	0	2/10	−0.085 02 − 0.057 24i	−0.005 18 − 0.090 53i
4	2/9	3/10	0.027 78+0.041 67i	−0.025 + 0.025i
6	3/9	4/10	−0.026 09 − 0.001 69i	0.030 18 − 0.015 53i
8	4/9	0	0.041 67	−0.025
10	0	0	−0.026 09 + 0.001 69i	0.030 18+0.015 53i
12	0	0	0.027 78−0.041 67i	−0.025 − 0.25i
14	0	0	−0.085 02 + 0.057 24i	−0.005 18 + 0.090 53i

u	$\varphi_{3,2}/64$	$f_2(u)$
0	0.015 63	0
2	−0.004 74 + 0.007 99i	0
4	−0.001 74 − 0.000 35i	2/90
6	−0.000 81 + 0.000 35i	7/90
8	−0.001 04	16/90
10	−0.000 81 − 0.000 35i	25/90
12	−0.001 74 + 0.000 35i	24/90
14	−0.004 74 − 0.007 99i	16/90

表 7-11 例 7.14 两年后盈余的分布函数

u	$\Pr(U_1^{**}+W_2=u)$	$\Pr(U_2^{**}=u)$
-3	2/90	0
-1	7/90	0
1	16/90	16/81
3	25/90	25/81
5	24/90	24/81
7	16/90	16/81

习题

7.1 已知年总赔付为 0, 5, 10, 15, 20 的概率分别为 0.4, 0.3, 0.15, 0.1, 0.05. 年初的保费为 6. 每年的利息为年初所有可用资金的 10%, 赔付发生在年底.

(a) 精确计算 $\widetilde{\psi}(2,3)$.

(b) 在考虑了保险费和利息后, 将所得的分布以 5 为跨度离散化, 计算 $\widetilde{\psi}(2,3)$.

7.2 采用 FFT 和以 5 为跨度的离散化方法重新计算习题 7.1.

第 8 章　连续时间破产模型

8.1　引　　言

本章将讨论盈余随时间连续变化的模型. 为了降低分析的难度, 首先考虑索赔次数服从 Poisson 分布的模型. 在离散时间情形, 我们可以直接导出相关的结果. 而在连续时间情形, 在某些特定的情况下也能得到精确的解析解, 但是在一般情况下只能推导出破产概率的上界和近似解. 本节将介绍 Poisson 过程和连续时间破产问题.

8.1.1　Poisson 过程

我们将讨论 Poisson 过程 $\{N_t; t \geqslant 0\}$ 的基本性质, 这里用这个过程表示某业务组合的索赔计数过程. 因此 N_t 表示 $(0, t]$ 内的索赔次数. 下面给出 Poisson 过程的正式定义.

定义 8.1　*索赔计数过程 $\{N_t; t \geqslant 0\}$ 称为强度参数为 λ 的Poisson 过程, 若满足以下三个条件:*

(1) $N_0 = 0$;

(2) *该过程具有平稳独立增量;*

(3) *在长度为 t 的区间内, 索赔次数服从均值为 λt 的 Poisson 分布, 即对所有的 $s, t > 0$ 都有*

$$\Pr(N_{t+s} - N_s = n) = \frac{(\lambda t)^n \mathrm{e}^{-\lambda t}}{n!}, \quad n = 0, 1, \cdots. \tag{8.1}$$

平稳增量意味着在一个固定区间内的索赔次数只依赖于区间的长度, 而与区间的位置无关, 例如, 不受时间趋势的影响. **独立增量**则说明, 在统计意义上, 任意区间内的索赔次数与之前的任意与其不相交的区间内的索赔次数相互独立. 平稳和独立同时成立, 则**独立的平稳增量**过程可以看作是以任一时刻为起始点重新开始的原过程. 事实上, 条件 (2) 的平稳性假设只是为了让定义的阐述更明确, 并不是必须的, 它可以由条件 (3) 推出.

Poisson 过程有一个重要的性质: 每两次索赔之间的时间间隔相互独立, 且服从相同的均值为 $1/\lambda$ 的指数分布. 令 W_j 为第 $(j-1)$ 次和第 j 次索赔的时间间隔, j=1, 2,$\cdots$, 其中, W_1 表示第一次索赔的发生时间, 从而有

$$\Pr(W_1 > t) = \Pr(N_t = 0) = \mathrm{e}^{-\lambda t},$$

因此, W_1 服从均值为 $1/\lambda$ 的指数分布. 根据独立增量性, 可得到下面的等式,

$$\begin{aligned}\Pr(W_2 > t|W_1 = s) &= \Pr(W_1 + W_2 > s + t|W_1 = s)\\ &= \Pr(N_{t+s} = 1|N_s = 1)\\ &= \Pr(N_{t+s} - N_s = 0|N_s = 1)\\ &= \Pr(N_{t+s} - N_s = 0).\end{aligned}$$

再由条件 (3), 得

$$\Pr(W_2 > t|W_1 = s) = \mathrm{e}^{-\lambda t}.$$

因为上式对所有的 s 成立, 所以有 $\Pr(W_2 > t) = \mathrm{e}^{-\lambda t}$, 且 W_2 独立于 W_1. 类似地, $W_3, W_4, W_5, \cdots$ 相互独立, 且服从相同的均值为 $1/\lambda$ 的指数分布.

最后指出, 由于指数分布的无记忆性, 从某个时间点 $t_0 \geqslant 0$ 开始到最近一次索赔之间的时间间隔同样服从均值为 $1/\lambda$ 的指数分布. 即如果第 n 次索赔发生在 t_0 前 s 个时间单位的时刻, 则下一次索赔发生在 t_0 后 t 个时间单位的概率为 $\Pr(W_{n+1} > t + s|W_{n+1} > s) = \mathrm{e}^{-\lambda t}$. 从上式中容易看出, 不论 s 和 n 取何值, 这一概率都是同一指数分布的生存函数.

8.1.2 连续时间的相关问题

我们将采用复合 Poisson 过程来刻画总索赔额. 下面给出正式定义.

定义 8.2 设索赔计数过程 $\{N_t; t \geqslant 0\}$ 服从强度参数为 λ 的 Poisson 过程. 个体损失 $\{X_1, X_2, \cdots\}$ 为相互独立且服从相同分布的正随机变量, 累积分布函数为 $F(x)$, 均值 $\mu < \infty$, 且均与 N_t 独立. 设 S_t 为 $(0, t]$ 内的总损失额, N_t=0 时有 S_t=0; N_t >0 时有 $S_t = \sum\limits_{j=1}^{N_t} X_j$, 其中 X_j 为第 j 次损失量. 可见, 对于固定的 t, S_t 服从复合 Poisson 分布. 称过程 $\{S_t; t \geqslant 0\}$ 为复合 Poisson 过程. 因为 $\{N_t; t \geqslant 0\}$ 具有独立平稳增量性, $\{S_t; t \geqslant 0\}$ 同样具有此性质, 且

$$\mathrm{E}(S_t) = \mathrm{E}(N_t)\mathrm{E}(X_j) = (\lambda t)(\mu) = \lambda\mu t.$$

假设保费连续收取, 单位时间内的保费收入为常数 c, 即 $(0, t]$ 内收取的净保费为 ct(为了数学上的简便, 此处不考虑利率). 进一步假设保费中含有正的附加保费, 也就是说 $ct > \mathrm{E}(S_t)$, 可简化为 $c > \lambda\mu$. 令

$$c = (1 + \theta)\lambda\mu, \tag{8.2}$$

其中, 称 $\theta > 0$ 为**相对附加安全系数或保费附加因子**.

上面已经详细说明了 4 个模型中的损失过程和保费收入过程, 接下来定义盈余过程

$$U_t = u + ct - S_t, \quad t \geqslant 0,$$

其中, $u = U_0$ 为初始盈余. 当 U_t 首次出现负值时, 称之为破产发生, 否则称之为生存状态. 从而得到无限时间生存概率的定义

$$\phi(u) = \Pr(U_t \geqslant 0 \text{ 对所有 } t \geqslant 0 | U_0 = u),$$

以及无限时间破产概率

$$\psi(u) = 1 - \phi(u).$$

我们的目的就是分析 $\phi(u)$ 和 $\psi(u)$.

8.2　调节系数和 Lundberg 不等式

本节我们将介绍一个特殊的量, 并通过它找出 $\psi(u)$ 的界. 因为只是一个界, 所以很容易得到, 而且作为上界它也提供了对破产概率的一种保守估计.

8.2.1　调节系数

很难从直观上给出定义调节系数的动机, 这里我们只是陈述定义. 在下面的讨论中, 一般用 X 表示索赔量随机变量.

定义 8.3　令 $t = \kappa$ 为下列等式的最小正值解

$$1 + (1 + \theta)\mu t = M_X(t), \tag{8.3}$$

其中, $M_X(t) = \mathrm{E}(\mathrm{e}^{tX})$ 是索赔量随机变量 X 的矩母函数. 如果上述值存在, 则称为调节系数.

下面考察定义 8.3 中解的存在性. 考虑 (t, y) 平面上的两条曲线: $y_1(t) = 1 + (1+\theta)\mu t$ 和 $y_2(t) = M_X(t) = \mathrm{E}(\mathrm{e}^{tX})$. $y_1(t)$ 是一条直线, 斜率为 $(1+\theta)\mu > 0$. 矩母函数可能并不存在或只对某些 t 存在. 这里, 我们假设对所有非负的 t, 矩母函数都存在. 因此有, $y_2'(t) = \mathrm{E}(X\mathrm{e}^{tX}) > 0$ 和 $y_2''(t) = \mathrm{E}(X^2\mathrm{e}^{tX}) > 0$. 因为 $y_1(0) = y_2(0) = 1$, 而 $y_2'(0) = \mathrm{E}(X) = \mu < (1 + \theta)\mu = y_1'(0)$, 所以这两条曲线在 $t = 0$ 有相同的起点, 且 $y_2(t)$ 一开始位于 $y_1(t)$ 的下面, 但因 $y_2'(t) > 0$ 和 $y_2''(t) > 0$, $y_2(t)$ 最终会在某一点 $\kappa > 0$ 处穿过 $y_1(t)$. κ 即为调节系数.

方程 (8.3) 可能并不存在正值解, 比如说, 个体索赔额分布的矩母函数不存在的情况. (例如, Pareto 分布、对数正态分布等).

例 8.4(指数分布的索赔量随机变量)　设 X 服从均值为 μ 的指数分布, 试确定调节系数.

解 指数分布的概率函数为 $F(x) = 1 - \mathrm{e}^{-x/\mu}, x \geqslant 0$. 矩母函数为 $M_X(t) = (1 - \mu t)^{-1}, t < \mu^{-1}$. 由方程 (8.3) 知, κ 满足下式

$$1 + (1+\theta)\mu\kappa = (1 - \mu\kappa)^{-1}. \tag{8.4}$$

如上文所述, κ=0 是方程的一个解, 方程的正值解为 $\kappa = \theta/[\mu(1+\theta)]$. 图 8-1 给出了 $\theta = 0.2$, $\mu = 0.1$ 时方程 (8.4) 左右两边的图像. 两个图像在 0 点和调节系数点 κ 相交, $\kappa = 0.2/1.2 = 0.166\ 7$. □

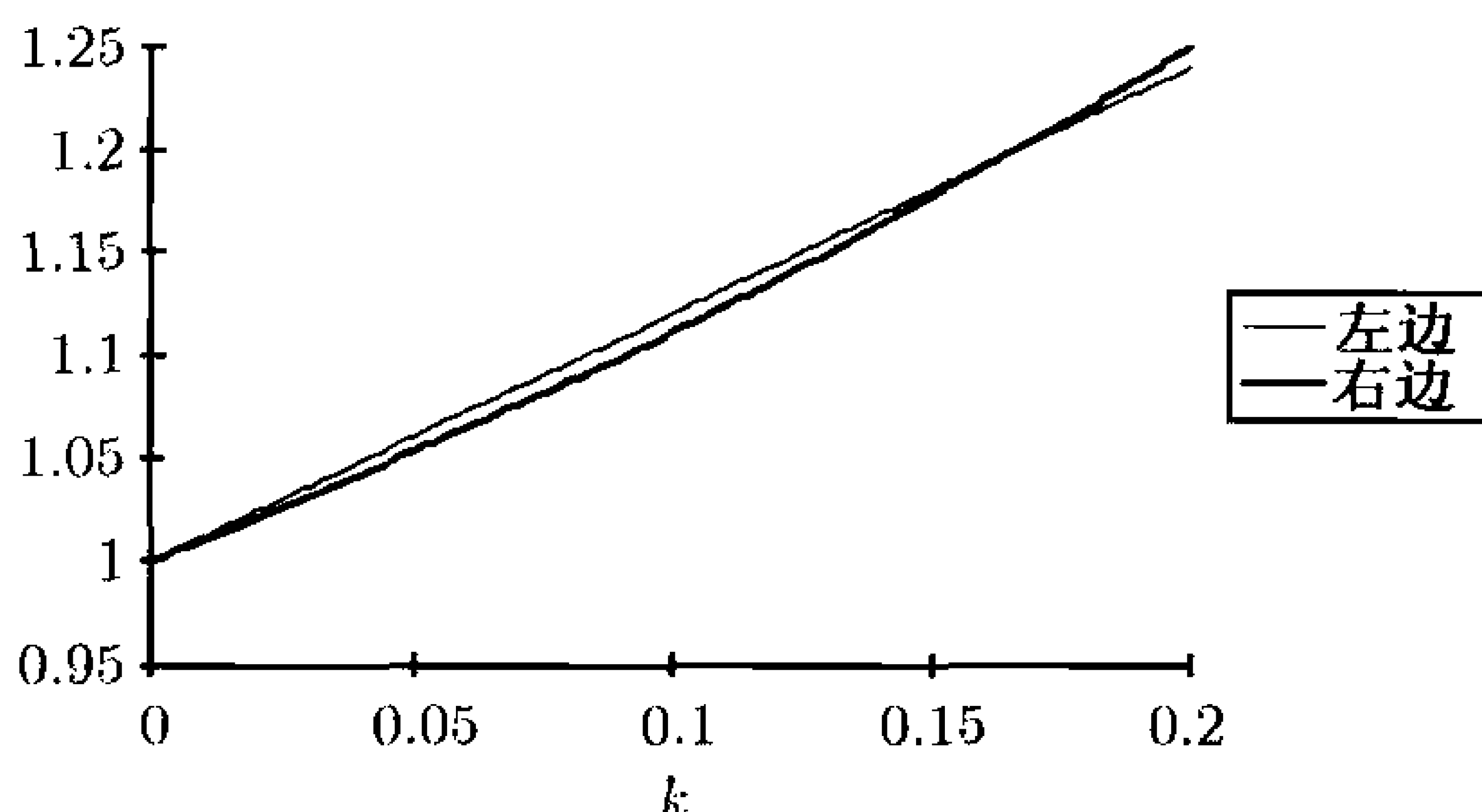


图 8-1 调节系数方程的左右边

例 8.5(gamma 分布) 设安全附加系数 $\theta = 2$, gamma 分布的参数 $\alpha = 2$. 用 β 表示 gamma 分布的尺度参数, 试求调节系数.

解 个体索赔额分布的密度函数为

$$f(x) = \beta^{-2} x \mathrm{e}^{-x/\beta}, \quad x > 0.$$

对于 gamma 分布, 其均值 $\mu = 2\beta$, 且

$$M_X(t) = \int_0^\infty \mathrm{e}^{tx} f(x)\mathrm{d}x = (1 - \beta t)^{-2}, \quad t < \frac{1}{\beta}.$$

根据方程 (8.3) 得

$$1 + 6\kappa\beta = (1 - \beta\kappa)^{-2},$$

整理上式得

$$6\beta^3\kappa^3 - 11\beta^2\kappa^2 + 4\beta\kappa = 0.$$

显然有因式分解

$$\kappa\beta(2\kappa\beta - 1)(3\kappa\beta - 4) = 0.$$

调节系数为原方程唯一的根[①], 即 $\kappa = 1/(2\beta)$. □

① 在因式分解式的两个正根中, 较大的根 $4/(3\beta)$ 是不合理的, 因为这时的矩母函数只有当小于 $1/\beta$ 时才存在. 在解这一类方程时, 调节系数总是等于最小的正根.

对于一般的个体索赔额分布, 很难像上面两例中那样精确地解出 κ 的值, 通常需要借助数值方法进行求解. 此时, 常需要给 κ 一个初始估计值. 注意, 从方程 (8.3) 可以得到

$$\begin{aligned}1+(1+\theta)\mu\kappa &= \mathrm{E}(\mathrm{e}^{\kappa X})\\ &= \mathrm{E}\left(1+\kappa X+\frac{1}{2}\kappa^2X^2+\cdots\right)\\ &> \mathrm{E}\left(1+\kappa X+\frac{1}{2}\kappa^2X^2\right)\\ &= 1+\kappa\mu+\frac{1}{2}\kappa^2\mathrm{E}(X^2).\end{aligned}$$

不等式两边同时减 $1+\kappa\mu$, 然后同时除以 κ, 结果如下

$$\kappa<\frac{2\theta\mu}{\mathrm{E}(X^2)}. \tag{8.5}$$

(8.5) 式的右端往往可以作为 κ 的初始估计值. 本节习题中将另外给出一些调节系数 κ 满足的不等式.

例 8.6　已知总损失随机变量的方差等于其均值的 3 倍. 求调节系数 κ 的界.

解　对于复合 Poisson 分布, $\mathrm{E}(S_t)=\lambda\mu t$, $\mathrm{Var}(S_t)=\lambda t\mathrm{E}(X^2)$, 所以从题意知 $\mathrm{E}(X^2)=3\mu$. 从而, 由 (8.5) 式, 有 $\kappa<2\theta/3$. □

定义

$$H(t)=1+(1+\theta)\mu t-M_X(t). \tag{8.6}$$

则调节系数 $\kappa>0$ 满足 $H(\kappa)=0$. 用 Newton-Raphson 公式求解该方程

$$\kappa_{j+1}=\kappa_j-\frac{H(\kappa_j)}{H'(\kappa_j)},$$

其中

$$H'(t)=(1+\theta)\mu-M'_X(t)$$

以某个 κ_0 作为初值逐步迭代. 因为 $H(0)=0$, 要注意避免迭代值趋向 0 的情况.

例 8.7　设 Poisson 参数 $\lambda=4$, 保费收入为 $c=7$. 个体损失量分布如下

$$\Pr(X=1)=0.6,\quad \Pr(X=2)=0.4.$$

求调节系数.

解　易求

$$\mu=E(X)=(1)(0.6)+(2)(0.4)=1.4,$$

$$\mathrm{E}(X^2)=(1)^2(0.6)+(2)^2(0.4)=2.2.$$

因而有, $\theta = c(\lambda\mu)^{-1} - 1 = 7(5.6)^{-1} - 1 = 0.25$. 从不等式 (8.5) 可以看出, κ 一定小于 $\kappa_0 = 2(0.25)(1.4)/2.2 = 0.318\ 2$. 现在有

$$M_X(t) = 0.6\mathrm{e}^t + 0.4\mathrm{e}^{2t},$$

根据 (8.6) 式知

$$H(t) = 1 + 1.75t - 0.6\mathrm{e}^t - 0.4\mathrm{e}^{2t}.$$

又有

$$M_X'(t) = (1\mathrm{e}^t)(0.6) + (2\mathrm{e}^{2t})(0.4),$$

所以

$$H'(t) = 1.75 - 0.6\mathrm{e}^t - 0.8\mathrm{e}^{2t}.$$

初始的估计值为 $\kappa_0 = 0.318\ 2$, 于是 $H(\kappa_0) = -0.023\ 81$ 和 $H'(\kappa_0) = -0.586\ 5$. 从而得到下一个估计值

$$\kappa_1 = 0.318\ 2 - \frac{-0.023\ 81}{-0.586\ 5} = 0.277\ 6.$$

接下来, 有 $H(0.277\ 6) = -0.003\ 091$ 和 $H'(0.277\ 6) = -0.435\ 8$, 以及

$$\kappa_2 = 0.277\ 6 - \frac{-0.003\ 091}{-0.435\ 8} = 0.270\ 5.$$

如此继续, 得到 $\kappa_3 = 0.270\ 3$, $\kappa_4 = 0.270\ 3$, 所以调节系数 $\kappa = 0.270\ 3$, 精确到小数点后四位. □

调节系数方程 (8.3) 也有其他的定义形式. 特别地, 下面是 κ 的一个等价定义:

$$1 + \theta = \int_0^\infty \mathrm{e}^{\kappa x} f_e(x)\mathrm{d}x, \tag{8.7}$$

其中

$$f_e(x) = \frac{1 - F(x)}{\mu}, \quad x > 0, \tag{8.8}$$

是 4.3.3 节中讨论到的均衡分布的概率密度函数.

下面验证定义式 (8.7) 等价于 (8.3) 式. 从习题 4.15 可知

$$\int_0^\infty \mathrm{e}^{\kappa x} f_e(x)\mathrm{d}x = \frac{M_X(\kappa) - 1}{\mu\kappa}.$$

将上式中的 $M_X(\kappa)$ 用 $1 + (1 + \theta)\mu\kappa$ 替代, 立即得到 (8.7) 式.

8.2.2 Lundberg 不等式

调节系数最重要的应用体现在以下结果中.

定理 8.8　设 $\kappa>0$ 满足方程 (8.3). 则破产概率 $\psi(u)$ 满足如下不等式

$$\psi(u)\leqslant \mathrm{e}^{-\kappa u},\quad u\geqslant 0. \tag{8.9}$$

证明　令 $\psi_n(u)$ 表示破产发生在第 n 次索赔或第 n 次索赔之前的概率, n=0, 1, 2, $\cdots$. 下面对 n 用归纳法, 来证明 $\psi_n(u)\leqslant \mathrm{e}^{-\kappa u}$. 显然, $\psi_0(u)=0\leqslant \mathrm{e}^{-\kappa u}$. 现假设 $\psi_n(u)\leqslant \mathrm{e}^{-\kappa u}$, 希望证明 $\psi_{n+1}(u)\leqslant \mathrm{e}^{-\kappa u}$. 对于 $\psi_{n+1}(u)$, 考虑第一次索赔的情形, 索赔的发生时间服从指数分布, 概率密度函数为 $\lambda\mathrm{e}^{-\lambda t}$. 如果索赔发生在 $t>0$ 时刻, 则可以用来支付的盈余为 $u+ct$. 因此, 若索赔额超过了 $u+ct$, 破产在第一次索赔时发生. 上述事件发生的概率为 $1-F(u+ct)$. 若索赔额 x 满足 $0\leqslant x\leqslant u+ct$, 则不发生破产, 且索赔后仍有 $u+ct-x$ 的盈余. 破产仍有可能发生在接下来的 n 次索赔中. 因为盈余过程具有独立平稳增量性, 所以这个概率等价于以 $u+ct-x$ 作为初始盈余的盈余过程在前 n 次索赔中发生破产的概率. 因此, 利用全概率公式, 得到下面的递归式①

$$\psi_{n+1}(u)=\int_0^{\infty}\left[1-F(u+ct)+\int_0^{u+ct}\psi_n(u+ct-x)\mathrm{d}F(x)\right]\lambda\mathrm{e}^{-\lambda t}\mathrm{d}t.$$

接着, 用归纳假设可以得到

$$\begin{aligned}\psi_{n+1}(u)&=\int_0^{\infty}\left[\int_{u+ct}^{\infty}\mathrm{d}F(x)+\int_0^{u+ct}\psi_u(u+ct-x)\mathrm{d}F(x)\right]\lambda\mathrm{e}^{-\lambda t}\mathrm{d}t\\&\leqslant\int_0^{\infty}\left[\int_{u+ct}^{\infty}\mathrm{e}^{-\kappa(u+ct-x)}\mathrm{d}F(x)\right.\\&\qquad\left.+\int_0^{u+ct}\mathrm{e}^{-\kappa(u+ct-x)}\mathrm{d}F(x)\right]\lambda\mathrm{e}^{-\lambda t}\mathrm{d}t,\end{aligned}$$

其中用到了 $x>u+ct$ 时 $-\kappa(u+ct-x)>0$ 的事实. 将两个积分合并, 得到

$$\begin{aligned}\psi_{n+1}(u)&\leqslant\int_0^{\infty}\left[\int_0^{\infty}\mathrm{e}^{-\kappa(u+ct-x)}\mathrm{d}F(x)\right]\lambda\mathrm{e}^{-\lambda t}\mathrm{d}t\\&=\lambda\mathrm{e}^{-\kappa u}\int_0^{\infty}\mathrm{e}^{-\kappa ct}\left[\int_0^{\infty}\mathrm{e}^{\kappa x}\mathrm{d}F(x)\right]\mathrm{e}^{-\lambda t}\mathrm{d}t\\&=\lambda\mathrm{e}^{-\kappa u}\int_0^{\infty}\mathrm{e}^{-(\lambda+\kappa c)t}[M_X(\kappa)]\mathrm{d}t\end{aligned}$$

① Stieltjes 积分符号 "$\mathrm{d}F(x)$" 可以看作是为了记号上的简便, 它包含了 X 离散、连续或者混合分布的情形. 在连续情形下, 用 $f(x)\mathrm{d}x$ 代替 $\mathrm{d}F(x)$, 就是我们所熟悉的 Riemann 积分模式. 在离散情形下, 积分符号只是一个记号, 实质就是求和, 包括对混合概率函数求和.

$$
\begin{aligned}
&= \lambda M_X(\kappa)\mathrm{e}^{-\kappa u}\int_0^{\infty}\mathrm{e}^{-(\lambda+\kappa c)t}\mathrm{d}t \\
&= \frac{\lambda M_X(\kappa)}{\lambda+\kappa c}\mathrm{e}^{-\kappa u}.
\end{aligned}
$$

根据 (8.3) 式和 (8.2) 式

$$\lambda M_X(\kappa) = \lambda[1+(1+\theta)\kappa\mu] = \lambda+\kappa(1+\theta)\lambda\mu = \lambda+\kappa c,$$

所以 $\psi_{n+1}(u) \leqslant \mathrm{e}^{-\kappa u}$. 从而, 对所有的 n 都有 $\psi_n(u) \leqslant \mathrm{e}^{-\kappa u}$, 则 $\psi(u) = \lim\limits_{n\to\infty}\psi_n(u) \leqslant \mathrm{e}^{-\kappa u}$. □

这个结果非常重要, 它可以用来检验不同水平的初始盈余 u 和保费附加系数 θ 之间的相互影响, 而这 2 个参数都是由保险公司控制的. 假设某保险公司的初始盈余为 u, 可以接受的破产概率水平为 α(例如, $\alpha = 0.01$). 则附加费率为

$$\theta = \frac{u\left\{\mathrm{E}\left[\exp\left(-\frac{\ln\alpha}{u}X\right)\right]-1\right\}}{-\mu\ln\alpha}-1.$$

它保证 (8.3) 式在 $\kappa = (-\ln\alpha)/u$ 时成立. 根据定理 8.8, 有 $\psi(u) \leqslant \mathrm{e}^{-\kappa u} = \mathrm{e}^{\ln\alpha} = \alpha$. 另外, 如果要求满足某个特定的附加费率 θ, 那么使破产概率不超过 α 的初始盈余 u 为

$$u = \frac{-\ln\alpha}{\kappa},$$

因为此时同样有 $\psi(u) \leqslant \mathrm{e}^{-\kappa u} = \mathrm{e}^{\ln\alpha} = \alpha$.

通过 (8.9), 我们还能得出

$$\psi(\infty) = \lim_{u\to\infty}\psi(u) = 0. \tag{8.10}$$

因为破产概率是非负的, 则

$$0 \leqslant \psi(u) \leqslant \mathrm{e}^{-\kappa u}, \tag{8.11}$$

所以

$$0 \leqslant \lim_{u\to\infty}\psi(u) \leqslant \lim_{u\to\infty}\mathrm{e}^{-\kappa u} = 0,$$

从而得到结果 (8.10) 式. 生存概率为

$$\phi(\infty) = 1. \tag{8.12}$$

习题

8.1 已知 $\theta = 0.32$, 个体索赔额分布同例 8.5. 求调节系数.

8.2 个体损失量的密度函数为 $f(x) = \sqrt{\beta/(\pi x)}\mathrm{e}^{-\beta x}, x > 0$. 试求调节系数.

8.3 已知 $c = 3$, $\lambda = 4$, 个体损失量的密度函数为 $f(x) = \mathrm{e}^{-2x} + \frac{3}{2}\mathrm{e}^{-3x}$, $x > 0$. 计算调节系数 (不要用数值迭代法).

8.4 已知 $c = 2.99$, $\lambda = 1$, 个体损失量分布如下: $\Pr(X = 1) = 0.2$, $\Pr(X = 2) = 0.3$, $\Pr(X = 3) = 0.5$. 试用 Newton-Raphson 迭代法计算调节系数.

8.5 用 Newton-Raphson 迭代法重新计算习题 8.3, 根据 (8.5) 式估计初值.

8.6 用 X 表示个体损失量随机变量, $\mathrm{E}(X^3)$ 已知. 证明调节系数 κ 满足

$$\kappa < \frac{-3\mathrm{E}(X^2) + \sqrt{9[\mathrm{E}(X^2)]^2 + 24\theta\mu\mathrm{E}(X^3)}}{2\mathrm{E}(X^3)}.$$

同时证明上述不等式右端严格小于 (8.5) 式的界, 即 $2\theta\mu/\mathrm{E}(X^2)$.

8.7 若 $g''(x) \geqslant 0$, 由 Jensen 不等式可知 $\mathrm{E}[g(Y)] \geqslant g[E(Y)]$. 同时, 根据 4.3.3 节有

$$\int_0^\infty x f_e(x)\mathrm{d}x = \frac{\mathrm{E}(X^2)}{2\mu},$$

其中 $f_e(x)$ 由 (8.8) 式定义.

(a) 用 (8.7) 式和上面的结果证明

$$\kappa \leqslant \frac{2\mu\ln(1+\theta)}{\mathrm{E}(X^2)}.$$

(b) 证明: 当 $\theta > 0$ 时, $\ln(1+\theta) < \theta$, 从而可知 (a) 中给出的不等式比 (8.5) 式更严格. 提示: 考虑 $h(\theta) = \theta - \ln(1+\theta)$, $\theta > 0$.

(c) 如果已知最大索赔量 m, 试证明 (8.7) 式等价于

$$1 + \theta = \int_0^m \mathrm{e}^{\kappa x} f_e(x)\mathrm{d}x.$$

同时证明上式右边满足

$$\int_0^m \mathrm{e}^{\kappa x} f_e(x)\mathrm{d}x \leqslant \mathrm{e}^{\kappa m},$$

从而有

$$\kappa \geqslant \frac{1}{m}\ln(1+\theta).$$

8.8 在 4.3.3 节中已经证明, 如果 $F(x)$ 的平均剩余寿命为递增的 [即 $F(x)$ 的损失率为递减的], 则有

$$\int_x^\infty f_e(y)\mathrm{d}y \geqslant 1 - F(x), \quad x \geqslant 0.$$

(a) 设 Y 的概率密度函数为 $f_e(y)$, $y \geqslant 0$, 及 X 的累积分布函数 $F(x)$. 证明

$$\Pr(Y > y) \geqslant \Pr(X > y), \quad y \geqslant 0,$$

以及因此得到

$$\Pr(e^{\kappa Y} > t) \geqslant \Pr(e^{\kappa X} > t), \quad t \geqslant 1.$$

(b) 用 (a) 证明 $E(e^{\kappa Y}) \geqslant E(e^{\kappa X})$.

(c) 用 (b) 证明 $\kappa \leqslant \theta/[\mu(1+\theta)]$.

(d) 如果题中所给不等式反向, 试证明

$$\kappa \geqslant \frac{\theta}{\mu(1+\theta)}.$$

8.9 假设 $\kappa > 0$ 满足方程 (8.3), 同时有

$$S(x) \leqslant \rho e^{-\kappa x} \int_0^\infty e^{\kappa y} dF(y), \tag{8.13}$$

其中, $0 < \rho \leqslant 1$, $S(x) = 1 - F(x)$. 证明: $\psi(u) \leqslant \rho e^{-\kappa u}$, $u \geqslant 0$. **提示**: 采用定理 8.8 的方法.

8.10 接上题. 用分部积分证明

$$\int_x^\infty e^{\kappa y} dF(y) = e^{\kappa x} S(x) + \kappa \int_x^\infty e^{\kappa y} S(y) dy, \quad x \geqslant 0.$$

8.11 假设 $F(x)$ 的风险率递减 (见 4.3.3 节), 试证 $S(y) \geqslant S(x)S(y-x)$, $x \geqslant 0$, $y \geqslant x$. 然后利用习题 8.10 证明 $\rho^{-1} = E(e^{\kappa X})$ 满足 (8.13) 式, 以及用 (8.3) 式得出以下结论

$$\psi(x) \leqslant [1 + (1+\theta)\kappa\mu]^{-1} e^{-\kappa x}, \quad x \geqslant 0.$$

8.12 设 $F(x)$ 的损失率为 $\mu(x) = -(d/dx) \ln S(x)$ 满足 $\mu(x) \leqslant m < \infty$, $x \geqslant 0$. 利用习题 8.10 的结果证明 $\rho = 1 - \kappa/m$ 满足 (8.13) 式且有

$$\psi(x) \leqslant (1 - \kappa/m) e^{-\kappa x}, \quad x \geqslant 0.$$

提示: 对于 $y > x$, 有 $S(y) \geqslant S(x) e^{-(y-x)m}$.

8.3 微积分方程

本节将给出破产概率 $\psi(u)$(或者生存概率 $\phi(u)$) 的一种显式表达式. 这将有助于接下来对更一般公式的推导.

定义 8.9 $G(u, y) = \Pr$(*初始盈余为 u 的条件下, 破产发生且在破产发生后瞬间的亏损量不超过 y*), $u \geqslant 0$, $y \geqslant 0$.

在上述定义的事件中, 破产后瞬间的盈余量处于 0 和 $-y$ 之间, 所以, 有

$$\psi(u) = \lim_{y\to\infty} G(u,y), \quad u \geqslant 0. \tag{8.14}$$

从而有以下结果.

定理 8.10　函数 $G(u,y)$ 满足下面的方程

$$\frac{\partial}{\partial u}G(u,y) = \frac{\lambda}{c}G(u,y) - \frac{\lambda}{c}\int_0^u G(u-x,y)\mathrm{d}F(x) - \frac{\lambda}{c}[F(u+y)-F(u)], \quad u \geqslant 0. \tag{8.15}$$

证明　再次考虑第一次索赔时的情况. 其发生时间服从指数分布, 概率密度函数为 $\lambda\mathrm{e}^{-\lambda t}$, 且在时刻 t 可用于支付索赔的盈余为 $u+ct$. 设索赔量为 x, 若 $0 \leqslant x \leqslant u+ct$, 则第一次索赔并不导致破产发生但盈余量减为 $u+ct-x$. 根据平稳独立增量性, 破产随后发生且亏损量不超过 y 的概率为 $G(u+ct-x,y)$. 另一种破产发生且亏损量不超过 y 的唯一可能是第一次索赔满足: $x > u+ct$ 且 $x \leqslant u+ct+y$, 因为若 $x > u+ct+y$, 则亏损量将超过 y. 索赔额满足 $u+ct < x \leqslant u+ct+y$ 的概率为 $F(u+ct+y) - F(u+ct)$. 从而根据全概率公式有

$$\begin{aligned} G(u,y) = \int_0^\infty \Bigg[\int_0^{u+ct} & G(u+ct-x,y)\mathrm{d}F(x) \\ & + F(u+ct+y) - F(u+ct)\Bigg]\lambda\mathrm{e}^{-\lambda t}\mathrm{d}t. \end{aligned}$$

为了便于将上面表达式对 u 求导, 做积分变量替换, 用 $z = u+ct$ 替换 t, 则有 $t = (z-u)/c$ 及 $\mathrm{d}t = \mathrm{d}z/c$. 变换后有

$$G(u,y) = \frac{\lambda}{c}\mathrm{e}^{(\lambda/c)u}\int_u^\infty \mathrm{e}^{-(\lambda/c)z}\left[\int_0^z G(z-x,y)\mathrm{d}F(x) + F(z+y) - F(z)\right]\mathrm{d}z.$$

根据微积分基本定理, 对函数 k 有 $\frac{\mathrm{d}}{\mathrm{d}u}\int_u^\infty k(z)\mathrm{d}z = -k(u)$. 据此对上式求导, 有

$$\begin{aligned} \frac{\partial}{\partial u}G(u,y) = \frac{\lambda}{c}G(u,y) + \frac{\lambda}{c}\mathrm{e}^{(\lambda/c)u}\Bigg\{ & -\mathrm{e}^{-(\lambda/c)u}\Bigg[\int_0^u G(u-x,y)\mathrm{d}F(x) \\ & + F(u+y) - F(u)\Bigg]\Bigg\}, \end{aligned}$$

即可得到定理结果. □

下面给出 $G(0,y)$ 的一个确切表达式.

定理 8.11　函数 $G(0,y)$ 有如下表达式

$$G(0,y) = \frac{\lambda}{c}\int_0^y [1-F(x)]\mathrm{d}x, \quad y \geqslant 0. \tag{8.16}$$

证明 首先, 有

$$0 \leqslant G(u,y) \leqslant \psi(u) \leqslant \mathrm{e}^{-\kappa u},$$

则

$$0 \leqslant G(\infty,y) = \lim_{u\to\infty} G(u,y) \leqslant \lim_{u\to\infty} \mathrm{e}^{-\kappa u} = 0,$$

所以 $G(\infty,y)=0$. 同时

$$\int_0^\infty G(u,y)\mathrm{d}u \leqslant \int_0^\infty \mathrm{e}^{-\kappa u}\mathrm{d}u = \kappa^{-1} < \infty.$$

令 $\tau(y)=\int_0^\infty G(u,y)\mathrm{d}u$, 可知 $0<\tau(y)<\infty$. 则将 (8.15) 式中的 u 从 0 到 ∞ 进行积分可得

$$-G(0,y) = \frac{\lambda}{c}\tau(y) - \frac{\lambda}{c}\int_0^\infty\int_0^u G(u-x,y)\mathrm{d}F(x)\mathrm{d}u - \frac{\lambda}{c}\int_0^\infty [F(u+y)-F(u)]\mathrm{d}u.$$

交换重积分的积分顺序, 有

$$\begin{aligned} G(0,y) = &-\frac{\lambda}{c}\tau(y) + \frac{\lambda}{c}\int_0^\infty\int_x^\infty G(u-x,y)\mathrm{d}u\mathrm{d}F(x) \\ &+ \frac{\lambda}{c}\int_0^\infty [F(u+y)-F(u)]\mathrm{d}u. \end{aligned}$$

接着对重积分的内部积分进行变量替换, 用 $v=u-x$ 替换 u, 有

$$\begin{aligned} G(0,y) &= -\frac{\lambda}{c}\tau(y) + \frac{\lambda}{c}\int_0^\infty\int_0^\infty G(v,y)\mathrm{d}v\mathrm{d}F(x) \\ &\quad + \frac{\lambda}{c}\int_0^\infty [F(u+y)-F(u)]\mathrm{d}u \\ &= -\frac{\lambda}{c}\tau(y) + \frac{\lambda}{c}\int_0^\infty \tau(y)\mathrm{d}F(x) + \frac{\lambda}{c}\int_0^\infty [F(u+y)-F(u)]\mathrm{d}u. \end{aligned}$$

由 $\int_0^\infty \mathrm{d}F(x)=1$, 上式右端的前两项可消去, 所以

$$\begin{aligned} G(0,y) &= \frac{\lambda}{c}\int_0^\infty [F(u+y)-F(u)]\mathrm{d}u \\ &= \frac{\lambda}{c}\int_0^\infty [1-F(u)]\mathrm{d}u - \frac{\lambda}{c}\int_0^\infty [1-F(u+y)]\mathrm{d}u. \end{aligned}$$

再次进行变量替换, 在第一个积分式中用 $x=u$ 替换 u, 第二个积分式中用 $x=u+y$ 替换 u, 可得

$$G(0,y) = \frac{\lambda}{c}\int_0^\infty [1-F(x)]\mathrm{d}x - \frac{\lambda}{c}\int_y^\infty [1-F(x)]\mathrm{d}x = \frac{\lambda}{c}\int_0^y [1-F(x)]\mathrm{d}x. \qquad \square$$

注意: 即使调节系数不存在, (8.16) 式仍然成立. 函数 $G(0,y)$ 本身也具有相当大的研究价值, 不过这里我们还是回到对 $\phi(u)$ 的分析.

定理 8.12 *无初始准备金的生存概率满足*

$$\phi(0)=\frac{\theta}{1+\theta}. \tag{8.17}$$

证明 由 $\mu=\int_0^\infty[1-F(x)]\mathrm{d}x$, 再根据 (8.16) 式, 有

$$\psi(0)=\lim_{y\to\infty}G(0,y)=\frac{\lambda}{c}\int_0^\infty[1-F(x)]\mathrm{d}x=\frac{\lambda\mu}{c}=\frac{1}{1+\theta}.$$

从而, $\phi(0)=1-\psi(0)=\theta/(1+\theta)$. □

$\phi(u)$ 的一般解可由下面的微积分方程得到, 同时满足由 (8.17) 式给出的初值条件.

定理 8.13 *最终生存概率 $\phi(u)$ 满足*

$$\phi'(u)=\frac{\lambda}{c}\phi(u)-\frac{\lambda}{c}\int_0^u\phi(u-x)\mathrm{d}F(x),\quad u\geqslant 0. \tag{8.18}$$

证明 在 (8.15) 式中令 $y\to\infty$, 并由 (8.14) 式, 得

$$\psi'(u)=\frac{\lambda}{c}\psi(u)-\frac{\lambda}{c}\int_0^u\psi(u-x)\mathrm{d}F(x)-\frac{\lambda}{c}[1-F(u)],\quad u\geqslant 0. \tag{8.19}$$

用生存概率 $\phi(u)=1-\psi(u)$, (8.19) 式可以表示为

$$\begin{aligned}-\phi'(u)&=\frac{\lambda}{c}[1-\phi(u)]-\frac{\lambda}{c}\int_0^u[1-\phi(u-x)]\mathrm{d}F(x)-\frac{\lambda}{c}[1-F(u)]\\&=-\frac{\lambda}{c}\phi(u)-\frac{\lambda}{c}\int_0^u\mathrm{d}F(x)+\frac{\lambda}{c}\int_0^u\phi(u-x)\mathrm{d}F(x)+\frac{\lambda}{c}F(u)\\&=-\frac{\lambda}{c}\phi(u)+\frac{\lambda}{c}\int_0^u\phi(u-x)\mathrm{d}F(x),\end{aligned}$$

其中用到 $F(u)=\int_0^u\mathrm{d}F(x)$. 定理得证. □

在选择使用 (8.18) 式还是 (8.19) 式时很大程度上取决于个人偏好. 一般使用 (8.18) 式, 因为它的数学表述更简便一些. 遗憾的是, 对于一般的 $F(x)$, 解的形式相当复杂, 所以我们将一般解的讨论推延到 8.4 节. 这里将给出 $F(x)$ 的一个特殊解.

例 8.14(指数分布) *假设同例 8.4 有 $F(x)=1-\mathrm{e}^{-x/\mu},x\geqslant 0$. 试求 $\phi(u)$.*

解 此时, (8.18) 式变成

$$\phi'(u)=\frac{\lambda}{c}\phi(u)-\frac{\lambda}{\mu c}\int_0^u\phi(u-x)\mathrm{e}^{-x/\mu}\mathrm{d}x.$$

用 $y=u-x$ 对 x 进行变量替换, 得到

$$\phi'(u)=\frac{\lambda}{c}\phi(u)-\frac{\lambda}{\mu c}\mathrm{e}^{-u/\mu}\int_0^u\phi(y)\mathrm{e}^{y/\mu}\mathrm{d}y. \tag{8.20}$$

希望能够消去 (8.20) 式中的积分项, 为此, 对上式关于 u 求导, 有

$$\phi''(u)=\frac{\lambda}{c}\phi'(u)+\frac{\lambda}{\mu^2c}\mathrm{e}^{-u/\mu}\int_0^u\phi(y)\mathrm{e}^{y/\mu}\mathrm{d}y-\frac{\lambda}{\mu c}\phi(u).$$

结合 (8.20) 式消去积分项

$$\phi''(u)=\frac{\lambda}{c}\phi'(u)-\frac{\lambda}{\mu c}\phi(u)+\frac{1}{\mu}\left[\frac{\lambda}{c}\phi(u)-\phi'(u)\right],$$

可以简化为

$$\phi''(u)=\left(\frac{\lambda}{c}-\frac{1}{\mu}\right)\phi'(u)=-\frac{\theta}{\mu(1+\theta)}\phi'(u).$$

两边同时乘以积分因子 $\mathrm{e}^{\theta u/[\mu(1+\theta)]}$, 表达式为

$$\frac{\mathrm{d}}{\mathrm{d}u}[\mathrm{e}^{\theta u/[\mu(1+\theta)]}\phi'(u)]=0.$$

对 u 积分, 得到

$$\mathrm{e}^{\theta u/[\mu(1+\theta)]}\phi'(u)=K_1.$$

在 (8.20) 式中令 $u=0$, 利用 (8.17) 式, 有

$$K_1=\phi'(0)=\frac{\lambda}{c}\frac{\theta}{1+\theta}=\frac{\lambda}{\lambda\mu(1+\theta)}\frac{\theta}{1+\theta}=\frac{\theta}{\mu(1+\theta)^2}.$$

从而, 有

$$\phi'(u)=\frac{\theta}{\mu(1+\theta)^2}\exp\left[-\frac{\theta u}{\mu(1+\theta)}\right],$$

对上式积分得到

$$\phi(u)=-\frac{1}{1+\theta}\exp\left[-\frac{\theta u}{\mu(1+\theta)}\right]+K_2.$$

由 (8.17) 式有 $\phi(0)=\theta/(1+\theta)$, 所以在 $u=0$ 时有 $K_2=1$. 因此

$$\phi(u)=1-\frac{1}{1+\theta}\exp\left[-\frac{\theta u}{\mu(1+\theta)}\right].$$

□

习题

8.13 假设索赔额服从指数分布, 概率密度函数同例 8.14, 即 $F(x)=1-\mathrm{e}^{-x/\mu}$.

(a) 由 (8.15) 式证明: $G(u,y)=\psi(u)F(y)$.

(b) 证明: 已知破产发生的条件下破产发生后瞬间的亏损量分布服从题中所给的指数分布.

8.14 此题为 $G(u,y)$ 和 $\psi(u)$ 的带瑕点的更新方程的推导. 这将有助于对 2 个函数推导各种性质.

(a) 对 (8.15) 式关于 u 从 0 到 t 积分, 用 (8.16) 式证明

$$
\begin{aligned}
G(t,y)=&\frac{\lambda}{c}\Lambda(t,y)-\frac{\lambda}{c}\int_0^t\Lambda(t-x,y)\mathrm{d}F(x)\\
&+\frac{\lambda}{c}\int_0^y[1-F(x)]\mathrm{d}x-\frac{\lambda}{c}\int_0^t[1-F(u)]\mathrm{d}u\\
&+\frac{\lambda}{c}\int_0^t[1-F(u+y)]\mathrm{d}u,
\end{aligned}
$$

其中, $\Lambda(x,y)=\int_0^x G(v,y)\mathrm{d}v$.

(b) 在 (a) 的结果中对 $\int_0^t\Lambda(t-x,y)\mathrm{d}F(x)$ 运用分部积分, 证明

$$
\begin{aligned}
G(t,y)=&\frac{\lambda}{c}\Lambda(t,y)-\frac{\lambda}{c}\int_0^t G(t-x,y)F(x)\mathrm{d}x\\
&+\frac{\lambda}{c}\int_0^{y+t}[1-F(x)]\mathrm{d}x-\frac{\lambda}{c}\int_0^t[1-F(u)]\mathrm{d}u.
\end{aligned}
$$

(c) 用 (b) 证明

$$
G(u,y)=\frac{\lambda}{c}\int_0^u G(u-x,y)[1-F(x)]\mathrm{d}x+\frac{\lambda}{c}\int_u^{y+u}[1-F(x)]\mathrm{d}x.
$$

(d) 证明

$$
\psi(u)=\frac{\lambda}{c}\int_0^u\psi(u-x)[1-F(x)]\mathrm{d}x+\frac{\lambda}{c}\int_u^{\infty}[1-F(x)]\mathrm{d}x.
$$

8.4 最大总损失

这里我们将推导微积分方程 (8.18) 式的一般解, 并满足由 (8.12) 式和 (8.17) 式给出的边界条件.

假设初始准备金为 u, 由于盈余过程具有平稳独立增量性, 则盈余量低于初始水平 u 的概率对于所有的 u 都相同, 而 $u=0$ 时概率为 $\psi(0)$, 则盈余量低于初始水平 u 的概率等于 $\psi(0)$.

这里的重要结论是, 当已知盈余低于初始水平 u 的事件发生时, 用随机变量 Y 表示盈余量首次跌至初始水平之下时的下跌量, 其均衡分布的概率密度函数为 $f_e(y)$, 其中 $f_e(y)$ 由 (8.8) 式给出.

定理 8.15 当已知盈余低于初始水平 u 的事件发生时, 随机变量 Y 表示这时的下跌量, 其概率密度函数为 $f_e(y)=[1-F(y)]/\mu$.

证明 函数 $G(u,y)$ 由定义 8.9 给出. 由于盈余过程具有平稳独立增量性, $G(0,y)$ 也表示了盈余量低于初始水平且下跌量不超过 y 的概率. 因此, 从定理 8.11 可知, 当已知盈余下跌的情况发生时, 下跌量的累积分布函数为

$$\begin{aligned}\Pr(Y\leqslant y)&=\frac{G(0,y)}{\psi(0)}\\&=\frac{\lambda}{c\psi(0)}\int_0^y[1-F(u)]\mathrm{d}u\\&=\frac{1}{\mu}\int_0^y[1-F(u)]\mathrm{d}u,\end{aligned}$$

求导即可得到所证结果. □

当下跌量为 y 时, 则发生下跌后瞬间的盈余量为 $u-y$. 根据盈余过程的平稳独立增量性, 若 $u-y>0$, 则此后破产发生的概率为 $\psi(u-y)$; 不然, 破产已经发生. 第二次跌破初始盈余的概率为 $\psi(0)$, 其下跌量同样有密度函数 $f_e(y)$, 且与第一次下跌独立. 由于 Poisson 过程的无记忆性, 每一次下跌后过程 "重新开始". 因此, 盈余量下跌的总次数 K 服从几何分布, 即 $\Pr(K=0)=1-\psi(0)$, $\Pr(K=1)=[1-\psi(0)]\psi(0)$. 更一般地, 有

$$\Pr(K=k)=[1-\psi(0)][\psi(0)]^k=\frac{\theta}{1+\theta}\left(\frac{1}{1+\theta}\right)^k,\quad k=0,1,2,\cdots,$$

上式由 $\psi(0)=1/(1+\theta)$ 可得. 这时的几何分布的参数 β(见附录 B) 为 $1/\theta$.

每次下跌发生后, 盈余立刻重新开始增长. 因此, 盈余量的最低值为 $u-L$, 其中 L 是所有下跌量的总和, 被称为**最大总损失**. 用 Y_j 表示第 j 次下跌量, 由于盈余过程的平稳独立增量性, $\{Y_1,Y_2,\cdots\}$ 是一系列独立同分布的随机变量 [密度函数均为 $f_e(y)$]. 下跌次数为 K, 则有

$$L=Y_1+Y_2+\cdots+Y_K,$$

若 $K=0$, 则 $L=0$. 因此, L 为复合几何随机变量, "索赔额密度" 为 $f_e(y)$.

显然, 初始盈余为 u 时, 若最大总损失 L 不超过 u, 则将一直生存下去, 即

$$\phi(u) = \Pr(L \leqslant u), \quad u \geqslant 0.$$

当 $y < 0$ 时, 令 $F_e^{*0}(y) = 0$; 当 $y \geqslant 0$ 时, $F_e^{*0}(y) = 1$. $F_e^{*k}(y) = \Pr\{Y_1 + Y_2 + \cdots + Y_k \leqslant y\}$ 表示 Y 和自身的 k 重卷积分布的累积分布函数. 从而得到 $\phi(u)$ 的一般表达式

$$\phi(u) = \sum_{k=0}^{\infty} \frac{\theta}{1+\theta} \left(\frac{1}{1+\theta}\right)^k F_e^{*k}(u), \quad u \geqslant 0.$$

按照破产概率的定义, 这个一般解也可以表示为

$$\psi(u) = \sum_{k=1}^{\infty} \frac{\theta}{1+\theta} \left(\frac{1}{1+\theta}\right)^k S_e^{*k}(u), \quad u \geqslant 0,$$

其中 $S_e^{*k}(y) = 1 - F_e^{*k}(y)$. 显然, $\psi(u)$ 是复合几何随机变量 L 相应的生存概率, 其解析解可以用 6.4 节中介绍的方法得到. 对于重要的 Erlang 混合索赔额概率密度函数① 以及 8.5 节中所介绍的索赔额分布, 得到解析解

$$f(x) = \sum_{k=1}^{r} q_k \frac{\beta^{-k} x^{k-1} \mathrm{e}^{-x/\beta}}{(k-1)!},$$

其中权重 $q_k > 0$, 加权和为 1(见习题 8.17). 8.5 节还有对其他索赔额分布的结论.

也可以运用第 6 章中介绍的各种技巧得到复合几何分布的累积分布函数, 进而使用数值方法计算破产概率.

例 8.16 设个体损失额服从 Pareto 分布, $\alpha = 3$, 均值为 500. 附加安全系数 $\theta = 0.2$. 计算 $\phi(u)$, $u = 100, 200, 300, \cdots$

解 首先要计算累积分布函数 $F_e(u)$. 这可以从它的概率密度函数得到

$$f_e(u) = \frac{1 - F(u)}{\mu} = \frac{1 - [1 - (\frac{1\,000}{1\,000+u})^3]}{500} = \frac{1}{500}\left(\frac{1\,000}{1\,000+u}\right)^3,$$

它恰好为 $\alpha = 2$、均值为 1 000 的 Pareto 分布的密度函数. 这个新的 Pareto 分布为某个复合几何分布的索赔额分布, 其参数 $\beta = 1/\theta = 5$. 可以用第 6 章中介绍的任一方法计算该复合几何分布. 运用递归公式, 对 Pareto 分布以跨度 5 进行保持均值不变的离散化处理. 将得到的离散概率相加即得到累积分布函数. 表 8-1 列出了计算结果. □

---

① 任何连续正概率的密度函数都可以用一个 Erlang 混合概率密度函数以任意精确度逼近, 具体见 Tijms[129], 第 163 页. Erlang 分布是形状参数 α 为整数的 gamma 分布.

表 8-1 损失为 Pareto 分布的生存概率

u	$\phi(u)$	u	$\phi(u)$
100	0.193	5 000	0.687
200	0.216	7 500	0.787
300	0.238	10 000	0.852
500	0.276	15 000	0.923
1 000	0.355	20 000	0.958
2 000	0.473	25 000	0.975
3 000	0.561		

习题

8.15 设索赔数服从 Poisson 过程, 个体索赔额服从均值为 100 的指数分布. 相对附加安全系数 $\theta = 0.1$. 运用本节所述方法计算 $\psi(1\ 000)$. 以 50 为跨度离散化指数分布. 将计算结果与精确的破产概率进行比较 (见例 8.14).

8.16 考虑例 8.5 中的问题, 令 $\beta = 50$. 试用本节的方法 (以跨度 1 考虑离散化递归公式) 近似 $\psi(200)$. 将计算结果与例 8.19 中得到的精确的破产概率进行比较.

8.17 设索赔额概率密度函数为

$$f(x) = \sum_{k=1}^{r} q_k \frac{\beta^{-k} x^{k-1} \mathrm{e}^{-x/\beta}}{(k-1)!}, \quad x > 0,$$

其中 $\sum_{k=1}^{r} q_k = 1$. 这是一个混合 gamma 密度.

(a) 证明

$$f_e(x) = \sum_{k=1}^{r} q_k^* \frac{\beta^{-k} x^{k-1} \mathrm{e}{-x/\beta}}{(k-1)!}, \quad x > 0,$$

其中

$$q_k^* = \frac{\sum_{j=k}^{r} q_j}{\sum_{j=1}^{r} j q_j}, \quad k = 1, 2, \cdots, r,$$

并证明 $\sum_{k=1}^{r} q_k^* = 1$.

(b) 定义

$$Q^*(z) = \sum_{k=1}^{r} q_k^* z^k.$$

运用习题 6.35 的结果证明

$$\psi(u) = \sum_{n=1}^{\infty} c_n \sum_{j=0}^{n-1} \frac{(u/\beta)^j \mathrm{e}^{-u/\beta}}{j!}, \quad u \geqslant 0,$$

其中

$$C(z) = \left\{1 - \frac{1}{\theta}[Q^*(z) - 1]\right\}^{-1}$$

是一个复合几何概率生成函数, 其概率值可用以下递归式计算:

$$c_k = \frac{1}{1+\theta}\sum_{j=1}^{k} q_j^* c_{k-j}, \quad k = 1, 2, \cdots,$$

初值 $c_0 = \theta(1+\theta)^{-1}$. 其中, 当 $j \neq 1, 2, \cdots, r$ 时, $q_j^* = 0$.

(c) 利用 (b) 证明

$$\psi(u) = \mathrm{e}^{-u/\beta}\sum_{j=0}^{\infty} \overline{C}_j \frac{(u/\beta)^j}{j!}, \quad u \geqslant 0,$$

其中 $\overline{C}_j = \sum\limits_{k=j+1}^{\infty} c_k$, $j = 0, 1, \cdots$. 再根据 (b) 证明可采用以下递归式计算 $\overline{C}_n$

$$\overline{C}_n = \frac{1}{1+\theta}\sum_{k=1}^{n} q_k^* \overline{C}_{n-k} + \frac{1}{1+\theta}\sum_{k=n+1}^{\infty} q_k^*, \quad n = 1, 2, \cdots,$$

初值 $\overline{C}_0 = (1+\theta)^{-1}$.

8.18　(a) 利用习题 8.14(c) 证明

$$G(u,y) = \frac{1}{1+\theta}\int_0^u G(u-x,y) f_e(x)\mathrm{d}x + \frac{1}{1+\theta}\int_u^{y+u} f_e(x)\mathrm{d}x,$$

其中 $G(u,y)$ 由 8.3 节定义, 再利用习题 8.14(d) 证明

$$\psi(u) = \frac{1}{1+\theta}\int_0^u \psi(u-x) f_e(x)\mathrm{d}x + \frac{1}{1+\theta}\int_u^{\infty} f_e(x)\mathrm{d}x,$$

其中 $f_e(x)$ 由 (8.8) 式给出.

(b) 用概率论方法直接证明 (a). *提示*: 以首次盈余下跌量为条件运用全概率公式.

8.5　Cramér 渐近破产公式和 Tijms 近似

本节将给出分析破产概率的另一个有价值的工具, 这里也用到了调节系数 κ. 下面定理中的结果被称为 Cramér 渐近破产公式. 记号 $a(x) \sim b(x)$, $x \to \infty$, 表示 $\lim\limits_{x\to\infty} a(x)/b(x) = 1$.

定理 8.17　*设 $\kappa > 0$ 使 (8.3) 式成立, 则破产概率满足*

$$\psi(u) \sim C\mathrm{e}^{-\kappa u}, \quad u \to \infty, \tag{8.21}$$

其中

$$C = \frac{\mu\theta}{M_X'(\kappa) - \mu(1+\theta)}, \tag{8.22}$$

$M_X(t) = \mathrm{E}(\mathrm{e}^{tX}) = \int_0^\infty \mathrm{e}^{tx}\mathrm{d}F(x)$ *是索赔额随机变量 X 的矩母函数.*

证明 此结果的证明十分复杂, 将利用重要的更新定理以及习题 8.14(d) 关于 $\psi(u)$ 的带瑕点的更新方程 (或习题 8.18(a) 中所给的等价形式). 有兴趣的读者可以参考 Rolski et.al.[113] 中 5.4.2 节. □

因此, 除了定理 8.8 的 Lundberg 不等式, 当 u 比较大时, 破产概率具有与指数函数相似的形式. 值得注意的是, 要使 Lundberg 不等式 (8.9) 式成立, (8.22) 式中的 C 必须满足 $C \leqslant 1$. 同时, 尽管 (8.21) 式是渐近近似, 但却是相当精确的, 即使对于不是很大的 u(特别是当附加安全系数 θ 本身比较小时). 在继续下面的分析之前, 我们先来介绍一个重要的例子.

例 8.18(指数分布) *若 $F(x) = 1 - \mathrm{e}^{-x/\mu}, x \geqslant 0$, 写出其渐近破产公式.*

解 在例 8.4 中可知其调节系数为 $\kappa = \theta/[\mu(1+\theta)]$ 以及 $M_X(t) = (1-\mu t)^{-1}$. 从而

$$M'_X(t) = \frac{\mathrm{d}}{\mathrm{d}t}(1-\mu t)^{-1} = \mu(1-\mu t)^{-2}.$$

同时, 有

$$M'_X(\kappa) = \mu(1-\mu\kappa)^{-2} = \mu[1-\theta(1+\theta)^{-1}]^{-2} = \mu(1+\theta)^2.$$

因此, 根据 (8.22) 式知

$$C = \frac{\mu\theta}{\mu(1+\theta)^2 - \mu(1+\theta)} = \frac{\theta}{(1+\theta)(1+\theta-1)} = \frac{1}{1+\theta}.$$

渐近公式 (8.21) 变为

$$\psi(u) \sim \frac{1}{1+\theta}\exp\left[-\frac{\theta u}{\mu(1+\theta)}\right], \quad u \to \infty.$$

这和例 8.14 得到的精确破产概率相一致. □

当 $F(x)$ 不是指数分布时, $\psi(u)$ 的精确解会比较复杂 (包括 8.4 节中的复合几何分布的一般解). 文献 Tijms[129] 第 271~272 页利用定理 8.17 中 Cramér 渐近破产公式在 u 很大时的精确性, 给出了一个简单的解析近似解. 其想法是在 (8.21) 式中加入一个指数项来提高对小数值 u 的精确性. 因此, Tijms 近似定义为

$$\psi_T(u) = \left(\frac{1}{1+\theta} - C\right)\mathrm{e}^{-u/\alpha} + C\mathrm{e}^{-\kappa u}, \quad u \geqslant 0, \tag{8.23}$$

其中, α 使 Tijms 近似的均值与最大总损失的复合几何分布的均值相一致. 如 4.3.3 节中所述, 盈余下跌量 (见 8.4 节定义) 的均值为 $\mathrm{E}(Y) = \int_0^\infty y f_e(y)\mathrm{d}y = \mathrm{E}(X^2)/2\mu$, 其中 $\mu = \mathrm{E}(X)$, X 是一般的索赔额随机变量. 类似地, 盈余的下跌次数 K 服从参

数为 $1/\theta$ 的几何分布, 所以从附录 B 知 $\mathrm{E}(K)=1/\theta$. 由于最大总损失 L 满足复合几何分布, 从 (6.6) 式可以得到它的均值为

$$\mathrm{E}(L)=\mathrm{E}(K)\mathrm{E}(Y)=\frac{\mathrm{E}(X^2)}{2\mu\theta}.$$

同时, $\psi(u)=\Pr(L>u)$, 根据 3.1 节 (3.9) 式可知, 当 $k=1$ 及 $u=\infty$ 时 $\mathrm{E}(L)=\int_0^\infty\psi(u)\mathrm{d}u$. 因此, 为了使 Tijms 近似保持均值一致, 我们用 $\psi_T(u)$ 代替积分中的 $\psi(u)$. 从而由 (8.23) 式得

$$\begin{aligned}\int_0^\infty\psi_T(u)\mathrm{d}u&=\left(\frac{1}{1+\theta}-C\right)\int_0^\infty \mathrm{e}^{-u/\alpha}\mathrm{d}u+C\int_0^\infty \mathrm{e}^{-\kappa u}\mathrm{d}u\\&=\alpha\left(\frac{1}{1+\theta}-C\right)+\frac{C}{\kappa},\end{aligned}$$

令上式和 $\mathrm{E}(L)$ 相等, 得到

$$\alpha\left(\frac{1}{1+\theta}-C\right)+\frac{C}{\kappa}=\frac{\mathrm{E}(X^2)}{2\mu\theta},$$

从中解出

$$\alpha=\frac{\mathrm{E}(X^2)/(2\mu\theta)-C/\kappa}{1/(1+\theta)-C}.\tag{8.24}$$

综上所述, (8.23) 式给出了破产概率的 Tijms 近似, 其中 α 满足 (8.24) 式.

除了提供一个有着良好性质的简单解析近似外, Tijms 近似 $\psi_T(u)$ 在某些情况下还能精确地再现 $\psi(u)$ 的真实值. (习题 8.21 将对此现象进行分析.) 特别地, 如果索赔额的概率密度函数有形式 $f(x)=p(\beta^{-1}\mathrm{e}^{-x/\beta})+(1-p)(\beta^{-2}x\mathrm{e}^{-x/\beta})$, $x\geqslant 0$, 其中 $0\leqslant p<1$(当 $p=1$ 时就是指数分布密度函数, 而这种情况下不必使用 Tijms 近似), 可以证明 $\psi_T(u)=\psi(u)$. 看下面的例子.

例 8.19(gamma 分布, 其形状参数① 为 2)　*如同例 8.5, 设 $\theta=2$, 个体索赔额的密度函数为 $f(x)=\beta^{-2}x\mathrm{e}^{-x/\beta}$, $x\geqslant 0$. 试求破产概率的 Tijms 近似.*

解　矩母函数为 $M_X(t)=(1-\beta t)^{-2}, t<1/\beta$, 从中可以得到 $M_X'(t)=2\beta(1-\beta t)^{-3}$, 以及 $\mu=M_X'(0)=2\beta$. 在例 8.5 中我们已知调节系数 $\kappa(\kappa>0)$ 满足 $1+(1+\theta)\kappa\mu=M_X(\kappa)$, 则在此例中有 $1+6\beta\kappa=(1-\beta\kappa)^{-2}$, 从而 $\kappa=1/(2\beta)$. 下面先求出 Cramér 渐近破产公式. 因为 $M_X'(\kappa)=M_X'[1/(2\beta)]=2\beta(1-\frac{1}{2})^{-3}=16\beta$, 所以由 (8.22) 式得到

① 对于 gamma 分布, 其形状参数是附录 A 中所给的 α, 注意不要和 Tijms 近似中的 α 值混淆.

$$C=\frac{(2\beta)(2)}{16\beta-(2\beta)(1+2)}=\frac{2}{5},$$

又根据 (8.21) 式, $\psi(u)\sim\frac{2}{5}\mathrm{e}^{-u/(2\beta)}, u\to\infty$. 下面我们考虑 Tijms 近似 (8.23) 式, 此例中为

$$\psi_T(u)=\left(\frac{1}{1+2}-\frac{2}{5}\right)\mathrm{e}^{-u/\alpha}+\frac{2}{5}\mathrm{e}^{-u/(2\beta)}=\frac{2}{5}\mathrm{e}^{-u/(2\beta)}-\frac{1}{15}\mathrm{e}^{-u/\alpha}.$$

下面只需求出 α. 易知, $M_X''(t)=6\beta^2(1-\beta t)^{-4}$, 由此可得 $\mathrm{E}(X^2)=M_X''(0)=6\beta^2$. 盈余下跌量的均值为 $\mathrm{E}(Y)=\mathrm{E}(X^2)/2\mu=6\beta^2/4\beta=3\beta/2$. 又因为盈余下跌次数的均值为 $\mathrm{E}(K)=1/\theta=\frac{1}{2}$, 所以最大总损失的均值为 $\mathrm{E}(L)=\mathrm{E}(K)\mathrm{E}(Y)=3\beta/4$, 且 α 必须满足 $\mathrm{E}(L)=\int_0^\infty\psi_T(u)\mathrm{d}u$, 或者等价地满足 (8.24) 式, 即 α 由下式给出[①]

$$\alpha=\frac{\frac{3\beta}{4}-\frac{2}{5}(2\beta)}{\frac{1}{1+2}-\frac{2}{5}}=\frac{3\beta}{4}.$$

Tijms 近似为

$$\psi_T(u)=\frac{2}{5}\mathrm{e}^{-u/(2\beta)}-\frac{1}{15}\mathrm{e}^{-4u/(3\beta)},\quad u\geqslant 0.$$

正如上面提到的, 此处 $\psi(u)=\psi_T(u)$. □

另一类使 Tijms 近似值与真实的破产概率相一致的索赔额概率密度函数具有如下形式, $f(x)=p(\beta_1^{-1}\mathrm{e}^{-x/\beta_1})+(1-p)(\beta_2\mathrm{e}^{-x/\beta_2}), x\geqslant 0$. 如果 $0<p<1$, 此分布为 2 个指数分布的混合分布, 而如果 $p=\beta_2/(\beta_2-\beta_1)$, 则为 2 个独立的指数随机变量和, 均值分别为 β_1 和 β_2 且 $\beta_1\neq\beta_2$. 下面举例说明.

例 8.20(混合指数分布) 设 $\theta=4/11$, 个体索赔额分布密度函数为 $f(x)=\mathrm{e}^{-3x}+10\mathrm{e}^{-5x}/3, x\geqslant 0$. 试求破产概率的 Tijms 近似.

解 首先矩母函数为

$$M_X(t)=\int_0^\infty \mathrm{e}^{tx}f(x)\mathrm{d}x=(3-t)^{-1}+\frac{10}{3}(5-t)^{-1}.$$

因此, $M_X'(t)=(3-t)^{-2}+(10/3)(5-t)^{-2}$, 则 $\mu=M_X'(0)=\frac{1}{9}+\frac{10}{75}=\frac{11}{45}$. 从等式 (8.3) 知调节系数 $\kappa>0$ 满足 $1+\frac{1}{3}\kappa=(3-\kappa)^{-1}+\frac{10}{3}(5-\kappa)^{-1}$. 两边都乘以 $3(3-\kappa)(5-\kappa)$ 得到

$$3(\kappa-3)(\kappa-5)+\kappa(\kappa-3)(\kappa-5)=3(5-\kappa)+10(3-\kappa).$$

① 从例 8.5 可以看到, $1/\alpha$ 是调节系数方程的另一个根, 这并不是巧合. 用这种方法计算 α 适用于任意的索赔额分布, 包括那些破产概率值无法用 Tijms 近似精确地再现的分布.

即

$$3(\kappa^2-8\kappa+15)+\kappa^3-8\kappa^2+15\kappa=45-13\kappa.$$

整理得

$$0=\kappa^3-5\kappa^2+4\kappa=\kappa(\kappa-1)(\kappa-4),$$

所以 $\kappa=1$(方程的最小正根).

接下来推导 Cramér 渐近破产公式. 因为 $M_X'(\kappa)=M_X'(1)=\frac{1}{4}+\frac{10}{3}\frac{1}{16}=\frac{11}{24}$, 利用等式 (8.22) 得到

$$C=\frac{(\frac{11}{45})(\frac{4}{11})}{\frac{11}{24}-(\frac{11}{45})(\frac{15}{11})}=\frac{32}{45},$$

从而 Cramér 渐近破产公式为 $\psi(u)\sim\frac{32}{45}\mathrm{e}^{-u}, u\to\infty$.

根据等式 (8.23), 有

$$\psi_T(u)=\left(\frac{1}{1+\frac{4}{11}}-\frac{32}{45}\right)\mathrm{e}^{-u/\alpha}+\frac{32}{45}\mathrm{e}^{-u}=\frac{1}{45}\mathrm{e}^{-u/\alpha}+\frac{32}{45}\mathrm{e}^{-u}.$$

为了确定 α 的值, 注意到 $M_X''(t)=2(3-t)^{-3}+\frac{20}{3}(5-t)^{-3}$, 从而有 $\mathrm{E}(X^2)=M_X''(0)=\frac{2}{27}+\frac{20}{3}(\frac{1}{125})=\frac{86}{675}$. 因此最大总损失的均值为

$$\mathrm{E}(L)=\frac{\mathrm{E}(X^2)}{2\mu\theta}=\frac{\frac{86}{675}}{2(\frac{11}{45})(\frac{4}{11})}=\frac{43}{60}.$$

进而由方程 (8.24) 得到

$$\alpha=\frac{\frac{43}{60}-\frac{32}{45}}{\frac{1}{1+\frac{4}{11}}-\frac{32}{45}}=\frac{1}{4},$$

所以 Tijms 近似 $\psi_T(u)=\frac{1}{45}\mathrm{e}^{-4u}+\frac{32}{45}\mathrm{e}^{-u}$. 同样有 $\psi(u)=\psi_T(u)$. □

从 (8.23) 不难看出, 当 $\kappa<1/\alpha$ 时, $\psi_T(u)\sim C\mathrm{e}^{-\kappa u}$, $u\to\infty$. 在这种情况下, 当 $u=0$ 或 $u\to\infty$ 时, $\psi_T(u)$ 和 $\psi(u)$ 相等, 同时 $\psi_T(u)$ 和复合几何的均值相符. 可以证明, 使 $u\to\infty$ 时 $\psi_T(u)$ 和 $\psi(u)$ 渐近相符的充分条件为, 非指数分布的索赔额分布函数 $F(x)$ 有着非增或非减的平均剩余寿命函数 [即 $F(x)$ 有着非增或非减的损失率, 具体讨论见 4.3.3 节]. 有趣的是, 在前一种情形下有 $\psi_T(u)>C\mathrm{e}^{-\kappa x}$, 而在后一种情形下则是 $\psi_T(u)<C\mathrm{e}^{-\kappa x}$. 这些事实的具体证明见 Willmot[137].

下面的例子说明了 Cramér 渐近公式和 Tijms 近似的精度.

例 8.21(形状参数为 3 的 gamma 分布) 假设索赔额分布服从均值为 1 的 gamma 分布, 密度函数为 $f(x)=27x^2\mathrm{e}^{-3x}/2, x\geqslant 0$. 当附加安全系数 θ 分别取 0.25, 1, 4, 初始盈余 u 分别取 0.10, 0.25, 0.75, 1 时, 计算破产概率的精确表达式及其 Cramér 渐近公式和 Tijms 近似.

解 矩母函数为 $M_X(t) = (1 - t/3)^{-3}$.

$\psi(u)$ 的精确值可以运用习题 8.17 中的运算法则得到, 即 $\psi(u) = \mathrm{e}^{-3u} \sum\limits_{k=j+1}^{\infty} \overline{C}_j (3u)^j/j!, u \geqslant 0$, 其中的 $\overline{C}_j$ 可用下面的递归公式计算

$$\overline{C}_j = \frac{1}{1+\theta}\sum_{k=1}^{j} q_k^* \overline{C}_{j-k} + \frac{1}{1+\theta}\sum_{k=j+1}^{\infty} q_k^*, \quad j = 1, 2, 3, \cdots,$$

其中 $\overline{C}_0 = 1/(1+\theta)$, $q_1^* = q_2^* = q_3^* = 1/3$, 且 $q_k^* = 0$, $k > 3$. 计算结果见表 8-2 中 "精确值" 一列.

Cramér 渐近破产概率由 (8.21) 式近似给出, 其中 κ 是 (8.3) 式的最小正根, 根据不同的 θ 可以用 8.2.1 节所介绍的 Newton-Raphson 公式迭代求得. 系数 C 由 (8.22) 式得到. 计算结果见表 8-2 中 "Cramér" 一列.

Tijms 近似值可以从 (8.23) 式求出, α 满足 (8.24) 式, 计算结果见表 8-2 中 "Tijms" 一列.

表 8-2 的值也可以在 Tijms[129] 第 272 页和 Willmot[137] 中找到. 表中的数值说明在本例的情形, Tijms 近似值是对真实值的一个精确估计, 特别是对比较小的 θ 成立. Cramér 渐近破产公式对于较小的 θ 和 u 同样有着很高的精度. 由于此处的 gamma 分布有着递增的损失率 (见例 4.17), Tijms 近似破产概率值总是小于 Cramér 渐近破产概率, 这从表中结果也能看出. 当 $u \to \infty$ 时, 破产概率的精确值、Tijms 近似和 Cramér 渐近都将趋于同一值, 不过即使是 $u = 1$, 这三个值也是相当吻合的. □

表 8-2 gamma 损失分布的破产概率

θ	u	Exact	Cramér	Tijms
0.25	0.10	0.783 4	0.807 6	0.784 4
	0.25	0.756 2	0.770 8	0.757 1
	0.50	0.707 4	0.713 1	0.707 4
	0.75	0.657 7	0.659 7	0.657 3
	1.00	0.609 7	0.610 3	0.609 3
1.00	0.10	0.474 4	0.533 2	0.476 4
	0.25	0.434 2	0.470 0	0.436 1
	0.50	0.366 4	0.380 9	0.366 5
	0.75	0.303 3	0.308 8	0.302 6
	1.00	0.248 4	0.250 2	0.247 6
4.00	0.10	0.183 9	0.265 4	0.185 9
	0.25	0.159 4	0.210 6	0.161 5
	0.50	0.120 9	0.143 2	0.121 2
	0.75	0.088 2	0.097 4	0.087 5
	1.00	0.062 6	0.066 3	0.061 8

习题

8.19　证明 (8.22) 式可以表示为

$$C=\frac{\theta}{\kappa E(Y\mathrm{e}^{\kappa Y})},$$

其中 Y 的概率密度函数为 $f_e(y)$. 从而证明对于习题 8.17 有

$$\psi(u)\sim\frac{\theta}{\kappa\beta\sum_{j=1}^{r}jq_j^*(1-\beta\kappa)^{-j-1}}\mathrm{e}^{-\kappa u},\quad u\to\infty,$$

其中 $\kappa>0$ 满足

$$1+\theta=Q^*[(1-\beta\kappa)^{-1}]=\sum_{j=1}^{r}q_j^*(1-\beta\kappa)^{-j}.$$

8.20　由 8.3 节给出函数 $G(u,y)$ 的定义. 从习题 8.14(c) 的结果可以看出 Cramér 渐近破产公式可以推广为

$$G(u,y)\sim C(y)\mathrm{e}^{-\kappa u},\quad u\to\infty,$$

其中

$$C(y)=\frac{\mu\kappa\int_0^\infty \mathrm{e}^{\kappa t}\int_y^{t+y}f_e(x)\mathrm{d}x\mathrm{d}t}{M_X'(\kappa)-\mu(1+\theta)}.$$

(a) 证明当 $y\to\infty$ 上式就是 Cramér 渐近破产公式.

(b) 利用习题 8.13 证明, 当索赔额服从指数分布 (分布函数为 $F(x)=1-\mathrm{e}^{-x/\mu}$) 时, 上面 $G(u,y)$ 的近似对所有的 u 均为等式.

8.21　假设破产概率有以下形式

$$\psi(u)=C_1\mathrm{e}^{-r_1u}+C_2\mathrm{e}^{-r_2u},\quad u\geqslant 0,$$

其中 $C_1\neq 0, C_2\neq 0$ 且 $0<r_1<r_2$(不失一般性).

(a) 确定附加安全系数 θ.　　**(b)** 求调节系数 κ.

(c) 证明：$0<C_1\leqslant 1$.　　**(d)** 求 Cramér 渐近破产公式.

(e) 证明 $\psi_T(u)=\psi(u)$, 其中 $\psi_T(u)$ 是破产概率的 Tijms 近似.

8.22　设 $\theta=\frac{4}{5}$, 索赔额密度函数为 $f(x)=(1+6x)\mathrm{e}^{-3x}, x\geqslant 0$.

(a) 试求 Cramér 渐近破产公式.　　**(b)** 确定破产概率 $\psi(u)$.

8.23　设 $\theta=\frac{3}{11}$, 索赔额密度函数为 $f(x)=2\mathrm{e}^{-4x}+\frac{7}{2}\mathrm{e}^{-7x}, x\geqslant 0$.

(a) 试求 Cramér 渐近破产公式.　　**(b)** 确定破产概率 $\psi(u)$.

8.24　设 $\theta=\frac{3}{5}$, 索赔额密度函数为 $f(x)=3\mathrm{e}^{-4x}+\frac{1}{2}\mathrm{e}^{-2x}, x\geqslant 0$.

(a) 试求 Cramér 渐近破产公式.　　**(b)** 确定破产概率 $\psi(u)$.

8.25　设 $\theta=\frac{7}{5}$, 索赔额密度函数为 2 个指数分布的卷积, $f(x)=\int_0^x 3\mathrm{e}^{-3(x-y)}2\mathrm{e}^{-2y}\mathrm{d}y, x\geqslant 0$.

(a) 试求 Cramér 渐近破产公式.　　**(b)** 确定破产概率 $\psi(u)$.

8.6 布朗运动风险过程

本节将讨论布朗运动 (Wiener 过程) 和盈余过程 $\{U_t; t \geqslant 0\}$ 的关系, 其中

$$U_t = u + ct - S_t, \quad t \geqslant 0. \tag{8.25}$$

而总损失过程 $\{S_t; t \geqslant 0\}$ 定义为

$$S_t = X_1 + X_2 + \cdots + X_{N_t}, \quad t \geqslant 0,$$

其中 $\{N_t; t \geqslant 0\}$ 是强度参数为 λ 的 Poisson 过程, 且当 $N_t = 0$ 时有 $S_t = 0$. 和本章前几节一样, 我们仍假设个体损失 $\{X_1, X_2, \cdots\}$ 为独立同分布的正值随机变量, 且矩母函数都存在. 盈余过程 $\{U_t; t \geqslant 0\}$ 以斜率 c 连续递增, c 为单位时间内的保费率, 而且过程在随机时间 $\{T_1, T_2, \cdots\}$ 发生向下的跳跃 $\{X_1, X_2, \cdots\}$, 过程的轨道如图 8-2 所示. 图中, $u = 20$, $c = 35$, $\lambda = 3$, 且 X 服从均值为 10 的指数分布.

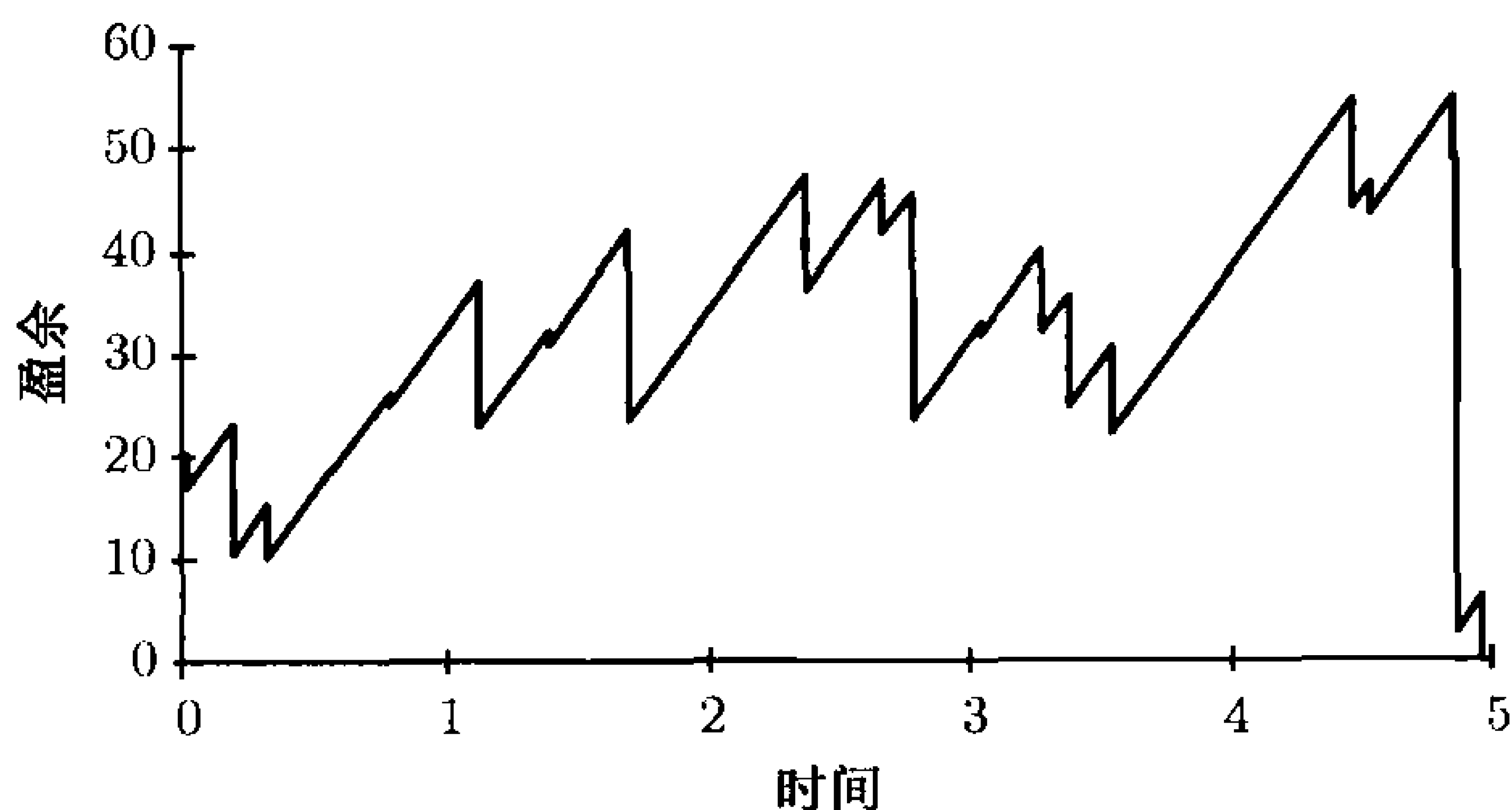


图 8-2 Poisson 盈余过程的样本轨道

令

$$Z_t = U_t - u = ct - S_t, \quad t \geqslant 0. \tag{8.26}$$

则 $Z_0 = 0$. 因为 S_t 为复合 Poisson 分布, 所以过程 $\{Z_t; t \geqslant 0\}$ 的均值为

$$\mathrm{E}(Z_t) = ct - \mathrm{E}(S_t) = ct - \lambda t \mathrm{E}(X),$$

方差为

$$\mathrm{Var}(Z_t) = \lambda t \mathrm{E}(X^2).$$

下面介绍基于布朗运动的随机过程.

定义 8.22 连续时间随机过程 $\{W_t; t \geqslant 0\}$ 称为布朗运动过程, 如果

(1) $W_0 = 0$;

(2) $\{W_t; t \geqslant 0\}$ 具有平稳独立增量性;

(3) 对所有的 $t>0$, W_t 服从均值为 0, 方差为 $\sigma^2 t$ 的正态分布.

布朗运动过程, 也称为**Wiener 过程**或**白噪声**, 已经广泛应用于描述各种自然现象. 当 $\sigma^2 = 1$ 时, 称为**标准布朗运动**. 1827 年, 英国植物学家 Robert Brown 发现了这个过程并用来描述液体或气体中粒子的无规则连续运动. 1905 年, Albert Einstein 也讨论过这个过程, 假设粒子和其所在媒介发生不间断的碰撞. Norbert Wiener 从 1918 年起在一系列的论文中对此过程提供了解析表达. 自此, 布朗运动开始应用于各个领域, 从量子力学到股票市场的资产定价, 已经成为现代金融理论的基础性模型.

定义 8.23 连续时间随机过程 $\{W_t; t \geqslant 0\}$ 称为有漂移的布朗运动过程, 若它具有布朗运动过程的性质, 而且 W_t 的均值为 μt, 其中 $\mu > 0$.

图 8-3 给出了一个有漂移的布朗运动的例子, 其中 $u=20, \mu=5, \sigma^2=600$. 图中过程的初始盈余为 20, 因此 W_t 的均值为 $20+5t$. $W_t - 20$ 是一个有漂移的布朗运动过程.

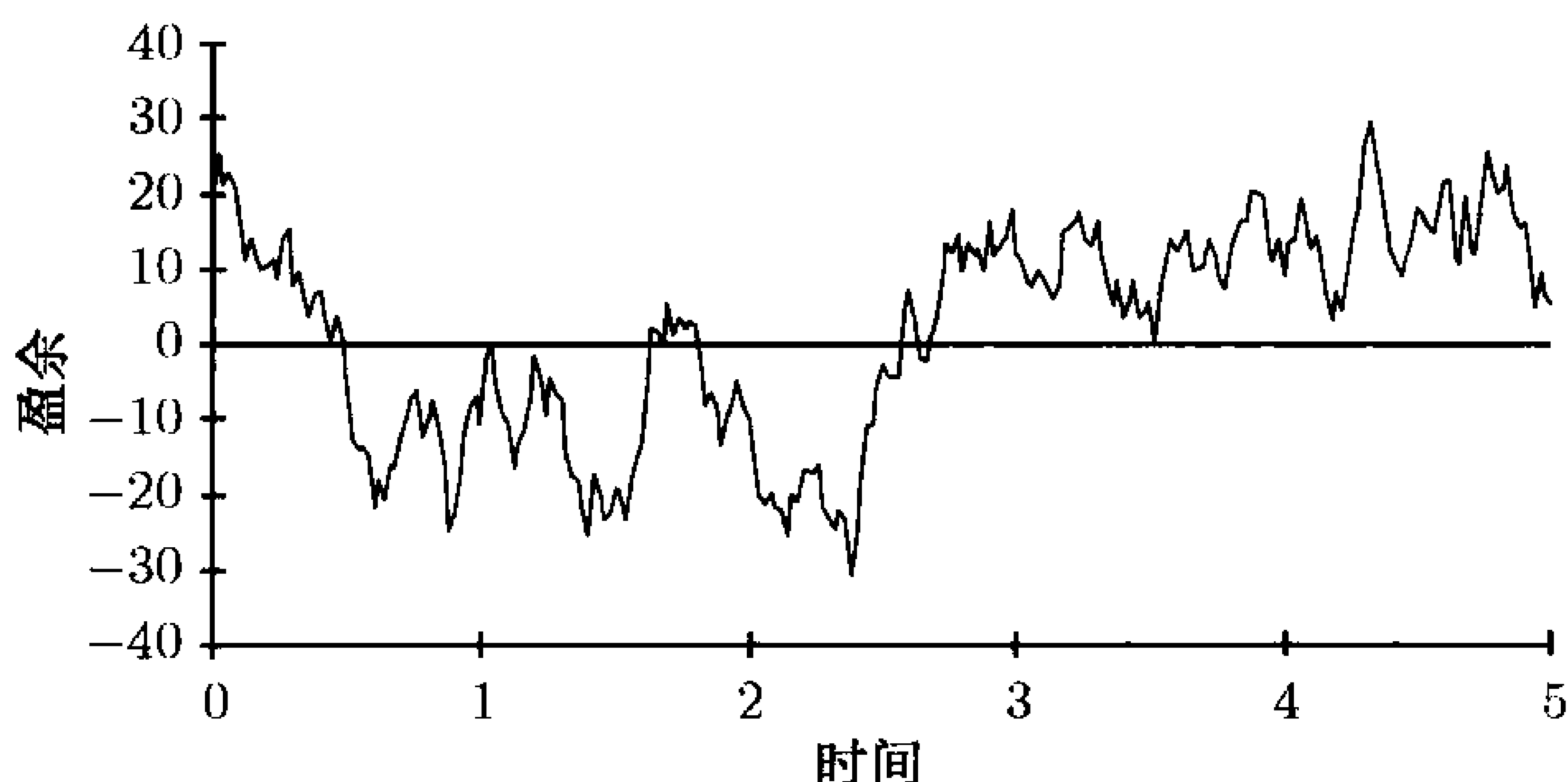


图 8-3 有漂移的布朗运动过程的样本轨道

下面将说明基于复合 Poisson 分布的风险过程的盈余过程 (8.26) 式是如何与有漂移的布朗运动过程相关联的. 对过程 (8.26) 式取极限状态, 令其向下跳跃的期望次数足够大, 同时跳跃幅度足够小. 由于有漂移的布朗运动过程表现出极小的均值 μ 和极小的方差 σ^2, 我们令两个过程的均值和方差函数相同. 这样, 有漂移的布朗运动过程可以认为是对基于复合 Poisson 的盈余过程的近似. 类似地, 也可以用复合 Poisson 过程近似布朗运动.

令

$$\mu = c - \lambda \mathrm{E}[X], \quad \sigma^2 = \lambda \mathrm{E}[X^2]$$

分别表示有漂移的布朗运动过程的均值和方差. 从而有

$$\lambda = \frac{\sigma^2}{\mathrm{E}[X^2]}, \tag{8.27}$$

$$c = \mu + \sigma^2 \frac{\mathrm{E}[X]}{\mathrm{E}[X^2]}. \tag{8.28}$$

为了体现极限过程, 将索赔额 X 视为另一个随机变量 Y 的尺度变换: $X = \alpha Y$. 其中, Y 具有固定的均值和方差. 则有

$$\lambda = \frac{\sigma^2}{\mathrm{E}[Y^2]} \cdot \frac{1}{\alpha^2},$$

$$c = \mu + \sigma^2 \frac{\mathrm{E}[Y]}{\mathrm{E}[Y^2]} \cdot \frac{1}{\alpha}.$$

因此, 为了使 $\lambda \to \infty$, 令 $\alpha \to 0$.

由于 $\{S_t; t \geqslant 0\}$ 是具有平稳独立增量的连续时间过程, 过程 $\{U_t; t \geqslant 0\}$ 和 $\{Z_t; t \geqslant 0\}$ 也具有相同的性质. 极限过程同样保持了该性质. 因为 $Z_0 = 0$, 按照定义 8.22 和定义 8.23, 只需证明对所有的 t, Z_t 在极限情形下服从均值为 μt, 方差为 $\sigma^2 t$ 的正态分布. 考虑 Z_t 的矩母函数

$$\begin{aligned} M_{Z_t}(r) &= M_{ct-S_t}(r) = \mathrm{E}\{\exp[r(ct - S_t)]\} \\ &= \exp(t\{rc + \lambda[M_X(-r) - 1]\}). \end{aligned}$$

从而

$$\begin{aligned} \frac{\ln M_{Z_t}(r)}{t} &= rc + \lambda[M_X(-r) - 1] = r[\mu + \lambda \mathrm{E}(X)] \\ &\quad + \lambda\left[1 - r\mathrm{E}(X) + \frac{r^2}{2!}\mathrm{E}(X^2) - \frac{r^3}{3!}\mathrm{E}(X^3) + \cdots - 1\right] \\ &= r\mu + \frac{r^2}{2}\lambda \mathrm{E}(X^2) - \lambda\left[\frac{r^3}{3!}\mathrm{E}(X^3) - \frac{r^4}{4!}\mathrm{E}(X^4) + \cdots\right] \\ &= r\mu + \frac{r^2}{2}\sigma^2 - \lambda\alpha^2\left[\alpha\frac{r^3}{3!}\mathrm{E}(Y^3) - \alpha^2\frac{r^4}{4!}\mathrm{E}(Y^4) + \cdots\right] \\ &= r\mu + \frac{r^2}{2}\sigma^2 - \sigma^2\left[\alpha\frac{r^3}{3!}\frac{\mathrm{E}(Y^3)}{\mathrm{E}(Y^2)} - \alpha^2\frac{r^4}{4!}\frac{\mathrm{E}(Y^4)}{\mathrm{E}(Y^2)} + \cdots\right]. \end{aligned}$$

因为除了 α, 其他项都是固定的, 则当 $\alpha \to 0$ 时, 有

$$\lim_{\alpha \to 0} M_{Z_t}(r) = \exp\left(r\mu t + \frac{r^2}{2}\sigma^2 t\right),$$

也就是均值为 μt、方差为 $\sigma^2 t$ 的正态分布的矩母函数, 这就证明了此极限过程是带漂移的布朗运动过程.

从图 8-2 可以清楚地看到, 过程 U_t 除了跳跃点外处处可导. 当跳跃点的数目无限增加时, 该过程将处处不可导. 布朗运动的另一个性质是它的轨道概率为 1 地为 t 的连续函数. 因为当 $\alpha \to 0$ 时, 跳跃幅度将变得非常小.

最后, 过程 U_t 在区间 $(0,t]$ 内经过的总距离为

$$D = ct + S_t = ct + X_1 + \cdots + X_{N_t},$$

期望值为

$$\begin{aligned}\mathrm{E}[D] = ct + \lambda t \mathrm{E}[X] &= t\left[\mu + \sigma^2 \frac{\mathrm{E}(Y)}{\mathrm{E}(Y^2)}\frac{1}{\alpha} + \sigma^2 \frac{\mathrm{E}(Y)}{\mathrm{E}(Y^2)}\frac{1}{\alpha}\right] \\ &= t\left[\mu + 2\sigma^2 \frac{\mathrm{E}(Y)}{\mathrm{E}(Y^2)}\frac{1}{\alpha}\right].\end{aligned}$$

当 $\alpha \to 0$ 时, 这个值将变得无穷大. 因此 $\lim_{\alpha\to 0} \mathrm{E}[D] = \infty$. 这意味着在一个有限区间内经过的距离期望值是无穷大! 关于布朗运动过程的性质, 更严格的讨论见 Karlin and Taylor[71] 编注[①]第 7 章.

因为 $Z_t = U_t - u$, 只需将 u 加到带漂移的布朗运动过程中, 即可使用 (8.27) 式和 (8.28) 式来得到过程 (8.26) 式的一个近似. 当然, λ 的值越大, 跳跃幅度越小, 近似的精度越高. 对于较大的保单组合 (例如, 整个公司), 这是非常适用的. 此时, 最终破产概率以及破产时间的分布可以很轻易地由带漂移的布朗运动过程近似得到. 8.7 节将会对此进行详细讨论. 类似地, 复合 Poisson 盈余过程也可以用来近似带漂移的布朗运动过程.

8.7 布朗运动和破产概率

设 $\{W_t; t \geqslant 0\}$ 是带漂移的布朗运动过程, 其均值为 μt, 方差为 $\sigma^2 t$. 用 $U_t = u + W_t$ 表示初始盈余 $U_0 = u$ 的带漂移的布朗运动过程.

考虑有限时间区间 $(0,\tau)$ 内的破产概率以及在破产发生条件下的破产时间的分布. 令 $T = \min_{t>0}\{t : U_t < 0\}$ 表示破产发生的时间 (若破产不发生, 则 $T = \infty$). 令 $\tau \to \infty$ 则得到最终破产概率.

在时刻 τ 之前破产发生的概率可以表示为

$$\begin{aligned}\psi(u,\tau) = 1 - \phi(u,\tau) = \Pr\{T < \tau\} &= \Pr\left\{\min_{0<t<\tau} U_t < 0\right\} \\ &= \Pr\left\{\min_{0<t<\tau} W_t < -U_0\right\} = \Pr\left\{\min_{0<t<\tau} W_t < -u\right\}.\end{aligned}$$

定理 8.24 对于上文描述的 U_t 过程, 破产概率可由下式给出

$$\psi(u,\tau) = \Phi\left(-\frac{u+\mu\tau}{\sqrt{\sigma^2\tau}}\right) + \exp\left(-\frac{2\mu}{\sigma^2}u\right)\Phi\left(-\frac{u-\mu\tau}{\sqrt{\sigma^2\tau}}\right), \tag{8.29}$$

---

① 中文版和英文影印版《随机过程初级教程 (第 2 版)》已由人民邮电出版社出版. —— 编者注

其中 $\Phi(\cdot)$ 是标准正态分布的累积分布函数.

证明 任意一条使 $U_\tau < 0$ 的 U_t 样本轨道必然在某个 $T < \tau$ 处穿过 $U_t = 0$. 对于这样的轨道 U_t, 我们定义一个新的轨道 U_t^*. 在 $t < T$ 时, U_t^* 和原来的轨道一样, 而在 $t > T$ 时, U_t^* 取原轨道关于 $U_t = 0$ 的反射. 即

$$U_t^* = \begin{cases} U_t, & t \leqslant T, \\ -U_t, & t > T. \end{cases} \tag{8.30}$$

反射轨道 U_t^* 的最终值满足 $U_\tau^* = -U_\tau$. 图 8-4 基于图 8-3 的样本轨道说明了上述反射过程.

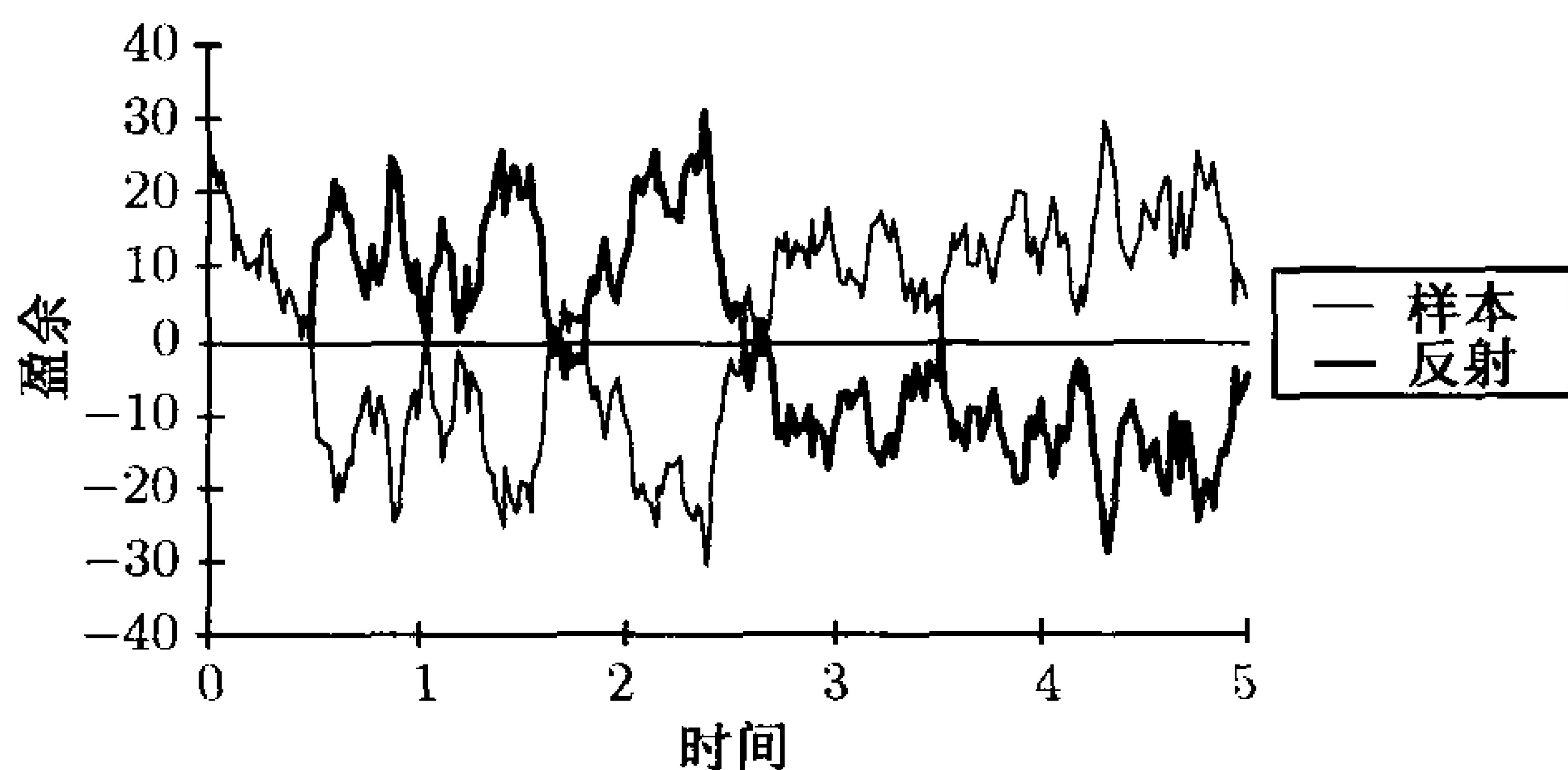


图 8-4 类型 B 的样本轨道及其反射轨道

现在考虑在 $(0, \tau)$ 内穿过 $U_t = 0$ 的所有轨道, 只有以下两种类型.

类型 A: 最终值 $U_\tau < 0$. 类型 B: 最终值 $U_\tau > 0$.

类型 B 中的任意一条轨道是类型 A 中某条轨道的反射. 因此, 样本轨道可以按照每个反射对进行考虑. 在 $(0, \tau)$ 内某个时间点发生破产的概率为上述所有反射对的总概率:

$$\psi(u, \tau) = \Pr\{T < \tau\} = \Pr\left\{\min_{0<t<\tau} U_t < 0\right\},$$

所有概率均在 $U_0 = u$ 的条件下发生. 这个概率是通过考虑类型 A 中所有满足 $U_\tau = x < 0$ 的原轨道得来的. 在此基础上增加所有相应的反射轨道 U_t^*, 则可以得到穿过破产界限的全部可能轨道. 注意, $U_\tau = 0$ 的情形并未考虑在内. 因为这个事件的概率为 0, 所以不必考虑. 图 8-4 中以正盈余为终点的样本轨道属于类型 B, 其反射轨道属于类型 A.

用 A_x 表示类型 A 中以 $U_\tau = x$ 为终点的所有可能的轨道集合, 用 B_x 表示类型 B 中以 $U_\tau = -x$ 为终点的所有可能的轨道集合. 令 $\Pr\{A_x\}$ 和 $\Pr\{B_x\}$ 分别表

示两个集合中的轨道的总概率[1]. 因此, 破产概率为

$$\Pr\{T<\tau\}=\int_{-\infty}^{0}\Pr\{U_\tau=x\}\frac{\Pr\{A_x\}+\Pr\{B_x\}}{\Pr\{U_\tau=x\}}\mathrm{d}x. \tag{8.31}$$

因为 A_x 表示以 x 为终点的所有可能的轨道集合, 所以

$$\Pr\{A_x\}=\Pr\{U_\tau=x\},$$

等式的右边是 U_τ 的概率密度函数. 从而有

$$\Pr\{T<\tau\}=\int_{-\infty}^{0}\Pr\{U_\tau=x\}\left[1+\frac{\Pr\{B_x\}}{\Pr\{A_x\}}\right]\mathrm{d}x. \tag{8.32}$$

由于 U_t-u 是带漂移的布朗运动过程, U_τ 服从均值为 $u+\mu\tau$, 方差为 $\sigma^2\tau$ 的正态分布, 所以

$$\Pr\{U_\tau=x\}=(2\pi\sigma^2\tau)^{-1/2}\exp\left[-\frac{(x-u-\mu\tau)^2}{2\sigma^2\tau}\right].$$

为了得到 $\Pr\{B_x\}/\Pr\{A_x\}$, 以所有可能的破产时间 T 为条件作条件概率, 则有

$$\begin{aligned}\frac{\Pr\{B_x\}}{\Pr\{A_x\}}&=\frac{\int_0^\tau\Pr\{B_x|T=t\}\Pr\{T=t\}\mathrm{d}t}{\int_0^\tau\Pr\{A_x|T=t\}\Pr\{T=t\}\mathrm{d}t}\\&=\frac{\int_0^\tau\Pr\{U_\tau=-x|T=t\}\Pr\{T=t\}\mathrm{d}t}{\int_0^\tau\Pr\{U_\tau=x|T=t\}\Pr\{T=t\}\mathrm{d}t}.\end{aligned}$$

因为 $T=t$ 意味着 $U_t=0$, 所以 $U_\tau|T=t$ 的条件概率密度函数和 $U_\tau-U_t$ 的概率密度函数相同. 过程 U_t 具有独立增量, 从而 $U_\tau-U_t$ 服从正态分布. 于是

$$\begin{aligned}\Pr\{U_\tau=x|T=t\}=\Pr\{U_\tau-U_t=x\}&=\frac{\exp\left\{-\frac{[x-\mu(\tau-t)]^2}{2\sigma^2(\tau-t)}\right\}}{\sqrt{2\pi\sigma^2(\tau-t)}}\\&=\frac{\exp\left\{-\frac{x^2-2x\mu(\tau-t)+\mu^2(\tau-t)^2}{2\sigma^2(\tau-t)}\right\}}{\sqrt{2\pi\sigma^2(\tau-t)}}\\&=\exp\left(\frac{x\mu}{\sigma^2}\right)\frac{\exp\left\{-\frac{x^2+\mu^2(\tau-t)^2}{2\sigma^2(\tau-t)}\right\}}{\sqrt{2\pi\sigma^2(\tau-t)}}.\end{aligned}$$

---

[1] 概率符号在此处的运用其实是错误的, 这些事件的真实概率为 0. 这里所说的概率实际上是概率密度, 对其积分可以得到具有正概率的集合概率.

类似地, 用 $-x$ 代替 x, 得到

$$\Pr\{U_\tau=-x|T=t\}=\frac{\exp(-\frac{x\mu}{\sigma^2})\exp\left\{-\frac{x^2+\mu^2(\tau-t)^2}{2\sigma^2(\tau-t)}\right\}}{\sqrt{2\pi\sigma^2(\tau-t)}}.$$

从而有

$$\begin{aligned}\frac{\Pr\{B_x\}}{\Pr\{A_x\}}&=\frac{\displaystyle\int_0^\tau\frac{\exp(-\frac{x\mu}{\sigma^2})\exp\left\{-\frac{x^2+\mu^2(\tau-t)^2}{2\sigma^2(\tau-t)}\right\}}{\sqrt{2\pi\sigma^2(\tau-t)}}\Pr\{T=t\}\mathrm{d}t}{\displaystyle\int_0^\tau\frac{\exp(\frac{x\mu}{\sigma^2})\exp\left\{-\frac{x^2+\mu^2(\tau-t)^2}{2\sigma^2(\tau-t)}\right\}}{\sqrt{2\pi\sigma^2(\tau-t)}}\Pr\{T=t\}\mathrm{d}t}\\&=\exp\left(-\frac{2\mu x}{\sigma^2}\right).\end{aligned}$$

根据 (8.31) 式, 最终得到

$$\begin{aligned}\psi(u,\tau)=\Pr\{T<\tau\}&=\int_{-\infty}^0\Pr\{U_\tau=x\}\left[1+\frac{\Pr\{B_x\}}{\Pr\{A_x\}}\right]\mathrm{d}x\\&=\Phi\left(-\frac{u-\mu\tau}{\sqrt{\sigma^2\tau}}\right)+\int_{-\infty}^0(2\pi\sigma^2\tau)^{-1/2}\exp\left[-\frac{(x-u-\mu\tau)^2}{2\sigma^2\tau}-\frac{2\mu x}{\sigma^2}\right]\mathrm{d}x\\&=\Phi\left(-\frac{u+\mu\tau}{\sqrt{\sigma^2\tau}}\right)+\int_{-\infty}^0(2\pi\sigma^2\tau)^{-1/2}\exp\left[-\frac{(x-u+\mu\tau)^2+4\mu u\tau}{2\sigma^2\tau}\right]\mathrm{d}x\\&=\Phi\left(-\frac{u+\mu\tau}{\sqrt{\sigma^2\tau}}\right)+\exp\left(-\frac{2\mu}{\sigma^2}u\right)\Phi\left(-\frac{u-\mu\tau}{\sqrt{\sigma^2\tau}}\right).\end{aligned}$$

□

推论 8.25 最终破产概率为

$$\psi(u)=1-\phi(u)=\Pr\{T<\infty\}=\exp\left(-\frac{2\mu}{\sigma^2}u\right).\tag{8.33}$$

在定理 8.24 中令 $\tau\to\infty$ 即可得到这个结果. 注意到 (8.29) 式中的分布是一个**带瑕点的分布**, 因为 $\tau\to\infty$ 时它的累积分布函数并不等于 1. 将带瑕点的分布以最终破产概率为条件计算条件概率就可得到正常的概率分布.

定理 8.26 已知破产发生时, 破产发生时间的分布为

$$\begin{aligned}\frac{\psi(u,\tau)}{\psi(u)}&=\Pr\{T<\tau|T<\infty\}\\&=\exp\left(\frac{2\mu}{\sigma^2}u\right)\Phi\left(-\frac{u+\mu\tau}{\sqrt{\sigma^2\tau}}\right)+\Phi\left(-\frac{u-\mu\tau}{\sqrt{\sigma^2\tau}}\right),\quad\tau>0.\end{aligned}\tag{8.34}$$

推论 8.27　破产发生时间的概率密度函数为

$$f_T(\tau) = \frac{u}{\sqrt{2\pi\sigma^2}}\tau^{-3/2}\exp\left[-\frac{(u-\mu\tau)^2}{2\sigma^2\tau}\right], \quad \tau > 0. \tag{8.35}$$

对 (8.34) 式关于 τ 求导即可得结果. 根据附录 A, 不难看出, 当 $\mu > 0$ 时, (8.35) 式是均值为 u/μ, 方差为 $u\sigma^2/\mu^3$ 的逆高斯分布的概率密度函数. $\mu = 0$ 时, 破产必然发生且破产时间 (8.35) 式的概率密度函数为

$$f_T(\tau) = \frac{u}{\sqrt{2\pi\sigma^2}}\tau^{-3/2}\exp\left(-\frac{u^2}{2\sigma^2\tau}\right), \quad \tau > 0,$$

其累积分布函数为

$$F_T(\tau) = 2\Phi\left(-\frac{u}{\sigma\tau^{1/2}}\right), \quad \tau > 0.$$

这个分布称为**指数为 1/2 的单边稳定率**.

上述结果可用来近似基于复合 Poisson 模型的盈余过程 (8.26) 式. 此时, $c = (1+\theta)\lambda\mathrm{E}(X)$, 其中 θ 是保费附加因子. 下面我们来推导这个近似, 先将 c 和 θ 代入 (8.27) 式和 (8.28) 式.

接着, 对于过程 (8.26), 由 (8.29) 式、(8.33) 式和 (8.35) 式得到

$$\begin{aligned}
\psi(u,\tau) &\doteq \Phi\left[\frac{u+\theta\lambda\tau\mathrm{E}(X)}{\sqrt{\lambda\tau\mathrm{E}(X^2)}}\right] \\
&\quad + \exp\left[-\frac{2\mathrm{E}(X)}{\mathrm{E}(X^2)}\theta u\right]\Phi\left[-\frac{u-\theta\lambda\tau\mathrm{E}(X)}{\sqrt{\lambda\tau\mathrm{E}(X^2)}}\right], \quad u>0, \tau>0, \\
\psi(u) &\doteq \exp\left[-\frac{2\mathrm{E}(X)}{\mathrm{E}(X^2)}\theta u\right], \qquad u > 0, \\
f_T(\tau) &\doteq \frac{u}{\sqrt{2\pi\lambda\mathrm{E}(X^2)}}\tau^{-3/2}\exp\left\{-\frac{[u-\theta\lambda\tau\mathrm{E}(X)]^2}{2\lambda\tau\mathrm{E}(X^2)}\right\}, \quad \tau > 0.
\end{aligned}$$

从而, 对于任意复合 Poisson 过程, 不难得到简单的数值近似. 例如, 假定破产发生, 破产时间的期望为

$$\mathrm{E}(T) = \frac{u}{\mu} = \frac{u}{\theta\lambda\mathrm{E}(X)}. \tag{8.36}$$

显然, 这个近似值的精度取决于各个量的相关程度.

注意, (8.36) 式所给的破产发生条件下破产时间的期望和 4 个用来描述盈余过程的关键量有关. 一个较高的初始盈余水平 (u) 会延迟破产时间, 而增大另外几个量将会减小破产时间的期望. 乍看起来似乎很令人惊讶, 但是并非如此, 例如增加附加保费将增大盈余的累积速度, 使破产不易发生. 因此, 若破产会发生, 那么它将在高附加保费带来高收益之前很快发生. 如果 λ 很大, 公司本身规模比较大, 事件

的发生会更快. 因此, 如果破产发生, 将会很快发生. 而若 $E(X)$ 的值比较大, 早期的索赔就可能消耗所有的初始盈余.

所有的这些只是直觉. 然而, 公式 (8.36) 还是揭示了各个因素是如何影响破产时间的期望的.

最后我们将指出, 用基于复合 Poisson 的风险过程 Z_t 来近似布朗运动过程也是可能的. 已知漂移系数和方差分别为 μ 和 σ^2, 由 (8.27) 式和 (8.28) 式可以得到

$$\mu = \theta\lambda E(X), \tag{8.37}$$

$$\sigma^2 = \lambda E(X^2). \tag{8.38}$$

为了简便起见, 固定跳跃幅度, 则 $E(X) = k$ 及 $E(X^2) = k^2$.

从而有

$$\lambda = \frac{\sigma^2}{k^2}, \tag{8.39}$$

$$\theta = \frac{\mu}{\lambda k} = \frac{\mu}{\sigma^2}k. \tag{8.40}$$

当 μ 和 σ^2 为固定值时, 通过选择 k 值来确定 λ 和 θ. 从而, 可用基于 Poisson 分布的过程来近似布朗运动, 其精度只依赖于参数 k. k 值越小, 跳跃幅度越小, 单位时间内的跳跃次数越大.

因此, 布朗运动过程的模拟可以通过基于 Poisson 分布的过程来完成. 在模拟 Poisson 过程时, 首先生成相继发生的事件的等待时间. 已知它们服从均值为 $1/\lambda$ 的指数分布, 所以当 k 变小时, λ 变大, 则平均等待时间将变小.

第三部分　经验模型的构造

第 9 章　数理统计基础

9.1　引　　言

在讨论经验模型和参数模型之前, 我们首先回顾数理统计的一些概念. 数理统计是一个非常广泛的专题, 它还包括很多本章未提及的内容. 对本章所涉及的内容, 我们假设读者已有所学习. 建立精算模型最重要的问题就是估计和假设检验. 在初等数理统计的教科书和课程中通常不涉及或只是简单介绍统计推断中的 Bayesian 方法. 本书的 12.4 节将对此方法作比较深入的介绍. Bayesian 方法也为第 16 章的信度理论提供了基础.

下面通过例子说明统计推断方法的作用. 假设你的老板要了解基本的牙医赔付的模型, 一种选择只是简单地给出模型, 你可以说这个模型为 $\mu = 5.123\ 9, \sigma = 1.034\ 5$ 的对数正态分布 (保留小数点后面很多位使给出的模型看起来比较真实). 当你的老板、监管者或律师将你置于证人的位置, 要求你回答如何得到这一结论时, 如果你的回答是 "我知道事情就是这样的", 这样的回答似乎是不充分的. 即使你声称有朋友在 Gamma Dental 公司也使用这个模型, 回答可能还是不充分的.

另一种选择是搜集有关的数据, 并利用这些数据来建立模型. 大多数模型由 2 部分组成: 一部分是分布函数的数学形式或者使用名称代表, 例如 Pareto 分布; 另一部分是参数, 它使模型具体化. 如果可以通过这种顺序将模型确定下来, 事情将变得很简单. 但在多数情况下, 我们需要先确定模型的参数, 然后才能决定是否采用这类模型.

由于参数估计是基于从总体中抽取的样本而不是整个总体, 所以得到的结果不是真实的参数值, 总会存在误差. 对这个问题的处理通常是通过区间估计表现的, 即给出某个取值的范围, 而不是仅仅给出具体的取值.

对于那些已知的参数分布我们将采用附录 A 和附录 B 中的参数化定义.

9.2　点　估　计

9.2.1　引言

无论怎样进行模型的估计, 都不可能使估计结果与真实的概率分布完全匹配. 而理想的情况是, 我们能够度量使用估计模型的误差, 但这点显然是不可能的! 如

果我们已知误差总量, 那么就可以按照这个量来调整估计结果, 从而消除任何的误差. 我们最多能够做到的是, 通过反复使用**这种方法**发现其内在误差, 而不是观察使用当前的估计模型产生了多少误差. 所以本节所讨论的是模型应用时的整体质量, 而不是某一特定应用时的效果.

这也是精算实践中很重要的一点. 最重要的就是选择合适的方法, 且让所有人了解: 即使是最好的模型, 也会由于未来的随机结果而产生不好的效果. 在北美精算师协会 (Society of Actuaries) 提出的精算原则草案 ([124], 779~780) 中关于人身风险责任准备金充足性部分 (也就是说, 公司有足够的资金用于支付保险合同的相关责任的概率), 很好地阐述了这个观点.

> 这里所示的充足性水平是面向未来的, 但精算模型通常是基于过去的经验得到的. 如果基于接下来发生的事件推断出模型的假设是不恰当的, 或是推断出所示的充足性水平过高或过低, 这样是不正确的.

建立模型时将会有各种误差, 有些误差这里不作介绍, 其中包括模型选择误差 (选择了错误的模型) 和抽样结构误差 (所抽取的样本与需要的对总体的推断不一致). 关于模型选择误差的例子是: 真实的分布情况为 Weibull 分布, 但模型选择了 Pareto 分布. 关于抽样结构误差的例子是: 用来自个人代理保单的索赔样本对网络销售的保单建立定价模型.

当利用抽取的样本得出有关整个总体的结论时, 若抽取的样本不能反映整体的情况就会产生误差, 而这个误差是我们能够度量的. 正如前面提到的, 我们不知道具体的一个样本是否能够代表整个总体. 但是我们能够估计一个不具代表性的样本对估计量的影响程度.

本节采用的办法是考虑总体的所有样本, 每一个样本都会产生一个估计值 (如概率、参数值或矩). 我们并不期望估计值总是与真实值相匹配. 对于一个明智的估计模型, 我们确实希望某些样本应使估计值与真实值完全匹配, 大部分样本使估计值与真实值接近, 只有很少的一些样本使估计值与真实值有较大偏差. 如果能够建立一个标准, 反映估计值与真实值匹配的程度, 我们就得到了判断估计程序质量的方法. 这里列出的方法通常被称为**经典的统计方法或频率学派方法**.

最后, 还需要说明**估计值**与**估计量**的区别, 前者表示按照某种估计方法对一组数据得到的具体计算结果, 后者表示产生估计值的规则或表达式. 估计值是一个数或函数值, 而估计量是一个随机变量或一个随机函数. 通常在具体运用这 2 个词时其含义将自然清楚地呈现出来.

9.2.2 估计量的评估

引言

有很多方法可用来度量估计量的质量, 这里介绍其中的 3 个方法, 并以具体例

子进行说明.

例 9.1 已知某总体包括 4 个等可能的值 1, 3, 5, 9. 有放回的随机抽取 2 个样本, 用样本均值估计总体均值.

例 9.2 已知某总体服从均值为 θ 的指数分布, 有放回的随机抽取 3 个样本, 用样本均值估计总体均值.

这两个例子明显是人造的, 因为在抽取样本前我们已经知道答案 (4.5 和 θ). 然而这一点在估计的过程中会产生明显的错误. 在实际应用中, 往往要在不知道估计量真实值的情况下估计误差.

无偏性

建立估计量时, 如果平均而言产生的误差能够彼此抵消, 则这个估计是很好的. 更正式的表述为, 设 θ 是要估计的量, $\hat{\theta}$ 是表示估计方法的随机变量, $\mathrm{E}(\hat{\theta}|\theta)$ 表示 θ 为真实参数值时估计量 $\hat{\theta}$ 的期望值.

定义 9.3 估计量 $\hat{\theta}$ 是无偏的, 如果 $\mathrm{E}(\hat{\theta}|\theta)=\theta$ 对所有 θ 成立. 且估计偏差定义为 $\mathrm{bias}_{\hat{\theta}}(\theta)=\mathrm{E}(\hat{\theta}|\theta)-\theta$.

估计偏差既依赖于估计量, 也可能依赖于 θ 的具体取值.

例 9.4 对例 9.1 计算将样本均值作为总体均值估计时的偏差.

解 总体均值是 $\theta=4.5$, 样本均值是 2 个观测值的平均, 这也是我们使用经验估计时的估计量. 在所有情况下我们都假设是随机抽取的. 也就是说, 任何样本量为 n 的样本都有相同的机会被抽到. 这样的抽取也意味着总体中的任何元素都有相同的机会被抽取. 例如, 这里就有 16 种等可能的方法得到样本, 它们被列举如下.

1,1	1,3	1,5	1,9	3,1	3,3	3,5	3,9
5,1	5,3	5,5	5,9	9,1	9,3	9,5	9,9

这产生了如下 16 种等可能的样本均值

1	2	3	5	2	3	4	6
3	4	5	7	5	6	7	9

将相同的值合并, 样本均值 (通常用 $\bar{X}$ 表示) 有如下的概率分布

x	1	2	3	4	5	6	7	9
$p_{\bar{X}}(x)$	1/16	2/16	3/16	2/16	3/16	2/16	2/16	1/16

估计量的期望值为

$$\mathrm{E}(\bar{X})=[1(1)+2(2)+3(3)+4(2)+5(3)+6(2)+7(2)+9(1)]/16=4.5,$$

所以在这个例子中样本均值是总体均值的无偏估计量. □

例 9.5 对例 9.2, 分别计算样本均值和样本中位数作为总体均值估计时的偏差.

解 样本均值为 $\bar{X} = (X_1 + X_2 + X_3)/3$, 其中 X_j 表示从指数总体中得到的一个观察, 它的期望值为

$$\begin{aligned}\mathrm{E}(\bar{X}) &= \mathrm{E}\left(\frac{X_1 + X_2 + X_3}{3}\right) = \frac{1}{3}[\mathrm{E}(X_1) + \mathrm{E}(X_2) + \mathrm{E}(X_3)] \\ &= \frac{1}{3}(\theta + \theta + \theta) = \theta,\end{aligned}$$

所以样本均值是总体均值的无偏估计.

样本中位数的考察会有些困难, 3 个观测值的中位数的分布函数可由下面的方法得到. 令 Y 表示我们关心的随机变量, 而 X 表示由总体观测的随机变量

$$\begin{aligned}F_Y(y) &= \Pr(Y \leqslant y) = \Pr(X_1, X_2, X_3 \leqslant y) + \Pr(X_1, X_2 \leqslant y, X_3 > y) \\ &\quad + \Pr(X_1, X_3 \leqslant y, X_2 > y) + \Pr(X_2, X_3 \leqslant y, X_1 > y) \\ &= F_X(y)^3 + 3F_X(y)^2[1 - F_X(y)] \\ &= [1 - \mathrm{e}^{-y/\theta}]^3 + 3[1 - \mathrm{e}^{-y/\theta}]^2\mathrm{e}^{-y/\theta}.\end{aligned}$$

密度函数为

$$f_Y(y) = F_Y'(y) = \frac{6}{\theta}(\mathrm{e}^{-2y/\theta} - \mathrm{e}^{-3y/\theta}).$$

该估计量的期望值为

$$\mathrm{E}(Y|\theta) = \int_0^\infty y\frac{6}{\theta}(\mathrm{e}^{-2y/\theta} - \mathrm{e}^{-3y/\theta})\mathrm{d}y = \frac{5\theta}{6}.$$

这个估计量显然是有偏的①, $\mathrm{bias}_Y(\theta) = 5\theta/6 - \theta = -\theta/6$. 平均而言这个估计量低估了真实值, 但只要再将样本中位数乘以 1.2 就可转化为无偏估计量. □

在例 9.2 中有 2 个估计量 (样本均值和 1.2 倍的样本中位数) 都是无偏的, 因此需要其他额外的标准来决定哪一个更好.

有些估计量表现出较小的偏差, 当样本量趋于无穷时, 偏差趋于零.

定义 9.6 设 $\hat{\theta}_n$ 是样本量为 n 时 θ 的估计量, 称估计量是渐近无偏的, 如果对任意的 θ 有

$$\lim_{n\to\infty} \mathrm{E}(\hat{\theta}_n|\theta) = \theta.$$

例 9.7 假设某随机变量服从区间 $(0, \theta)$ 上的均匀分布, 考虑估计量 $\hat{\theta}_n = \max(X_1, \cdots, X_n)$, 证明这个估计量是渐近无偏的.

① 样本中位数不可能是总体均值的一个较好的估计量, 研究这个例子是为了进行比较. 因为总体中位数是 $\theta\ln 2$, 样本中位数相对总体中位数也是有偏差的.

解 令 Y_n 表示 n 个样本中的最大者, 则

$$F_{Y_n}(y) = \Pr(Y_n \leqslant y) = \Pr(X_1 \leqslant y, \cdots, X_n \leqslant y) = [F_X(y)]^n = (y/\theta)^n,$$
$$f_{Y_n}(y) = \frac{ny^{n-1}}{\theta^n},\ 0 < y < \theta.$$

期望值为

$$\mathrm{E}(Y_n|\theta) = \int_0^\theta ny^n\theta^{-n}\mathrm{d}y = \frac{n}{n+1}y^{n+1}\theta^{-n}\bigg|_0^\theta = \frac{n\theta}{n+1}.$$

当 $n \to \infty$ 时, 极限为 θ, 所以这个估计量是渐进无偏的.

相合性

关于估计量的第二个可取的性质是当样本量很大时表现得比较好. 稍正式一些的描述为, 当样本量趋于无穷时, 估计量的误差大于一个小量的概率趋于零. 正式的定义如下.

定义 9.8 称估计量是相合的(这里通常称为弱相合的), 如果对所有 $\delta > 0$ 和任一个 θ, 有

$$\lim_{n\to\infty} \Pr(|\hat{\theta}_n - \theta| > \delta) = 0.$$

弱相合性的一个充分条件 (不是必要的) 为, 估计量是渐进无偏的并且 $\mathrm{Var}(\hat{\theta}_n) \to 0$.

例 9.9 证明, 如果随机变量的方差有限, 则样本均值是总体均值的一个相合估计.

解 由习题 9.2 知样本均值是无偏的, 另外,

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_{j=1}^n X_j\right) = \frac{1}{n^2}\sum_{j=1}^n \mathrm{Var}(X_j) = \frac{\mathrm{Var}(X)}{n} \to 0.$$

第 2 步计算结果是因为假设观察值独立而得到的. □

例 9.10 对于区间 $(0,\theta)$ 上的均匀分布, 证明其最大观测值是 θ 的相合估计.

解 由例 9.7 知, 最大观测值是渐进无偏的, 二阶矩为

$$\mathrm{E}(Y_n^2) = \int_0^\theta ny^{n+1}\theta^{-n}\mathrm{d}y = \frac{n}{n+2}y^{n+2}\theta^{-n}\bigg|_0^\theta = \frac{n\theta^2}{n+2},$$

所以

$$\mathrm{Var}(Y_n) = \frac{n\theta^2}{n+2} - \left(\frac{n\theta}{n+1}\right)^2 = \frac{n\theta^2}{(n+2)(n+1)^2} \to 0.$$ □

均方误差

相合性是很好的性质, 大部分估计量都有这个性质. 但真正有意义的是这个估计量的收敛不仅在平均意义上是正确的, 还要在大多数情况下都非常接近, 特别地, 比其他任何估计量都更接近. 受到相合性定义的启发产生了关于有限样本的估计标准, 估计量的质量可以用它与真实值的差距不超过 δ 的概率来度量, 也就是用 $\Pr(|\hat{\theta}_n - \theta| < \delta)$ 度量. 但 δ 的选择是任意的, 而人们更喜欢保持不变的度量标准, 所以要考虑 $\mathrm{E}(|\hat{\theta}_n - \theta|)$ 一平均的绝对误差. 但绝对值总会带来数学处理的不方便, 所以如下定义成为广泛接受的判断估计精度的标准.

定义 9.11　*估计量的均方误差 (MSE) 为*

$$\mathrm{MSE}_{\hat{\theta}}(\theta) = \mathrm{E}[(\hat{\theta} - \theta)^2|\theta].$$

注意, MSE 是参数真实值的函数, 一个估计量可能在参数的某些取值上表现得很好, 但在其他值上表现得很差.

例 9.12　*考虑未知参数 θ 的估计量 $\hat{\theta} = 5$, MSE 为 $(5-\theta)^2$, 当 θ 接近 5 时 MSE 很小, 但对于其他值 MSE 就变得很差. 当然这个估计是有偏的、非相合的, 除非 θ 恰好等于 5.*

根据一些定义可直接得到如下结果:

$$\mathrm{MSE}_{\hat{\theta}}(\theta) = \mathrm{E}\{[\hat{\theta} - \mathrm{E}(\hat{\theta}|\theta) + \mathrm{E}(\hat{\theta}|\theta) - \theta]^2|\theta\} = \mathrm{Var}(\hat{\theta}|\theta) + [\mathrm{bigs}_{\hat{\theta}}(\theta)]^2. \qquad (9.1)$$

如果只限于无偏估计量, 则上面所述估计量中最好的估计如下定义.

定义 9.13　*估计量 $\hat{\theta}$ 称为一致最小方差无偏估计 (UMVUE), 如果它是无偏的, 并且对于 θ 的任何真实值, 它具有比其他估计量更小的方差.*

因为只关心无偏估计, 所以上述定义与利用 MSE 定义有相同的效果. 我们也可以通过寻找关于 MSE 一致最优的估计量来推广这个定义, 但前面的例子已经指出这是不可行的. 虽然有一些现成的定理能够帮助人们确定 UMVUE. 然而, 确定这样的估计量是很困难的. 而另一方面, 在比较两个备选的估计量时, MSE 却是一个很有用的判别标准.

例 9.14　*在例 9.2 中比较样本均值和 1.2 倍样本中位数的 MSE.*

解　样本均值的方差为

$$\frac{\mathrm{Var}(X)}{3} = \frac{\theta^2}{3}.$$

乘以 1.2 后, 样本中位数的二阶矩为

$$\mathrm{E}[(1.2Y)^2] = 1.44\int_0^\infty y^2\frac{6}{\theta}(\mathrm{e}^{-2y/\theta} - \mathrm{e}^{-3y/\theta})\mathrm{d}y$$

$$= 1.44\frac{6}{\theta}\left[y^2\left(\frac{-\theta}{2}\mathrm{e}^{-2y/\theta} + \frac{\theta}{3}\mathrm{e}^{-3y/\theta}\right)\right.$$
$$- 2y\left(\frac{\theta^2}{4}\mathrm{e}^{-2y/\theta} - \frac{\theta^2}{9}\mathrm{e}^{-3y/\theta}\right)$$
$$\left.+ 2\left(\frac{-\theta^3}{8}\mathrm{e}^{-2y/\theta} + \frac{\theta^3}{27}\mathrm{e}^{-3y/\theta}\right)\right]\Bigg|_0^\infty$$
$$= \frac{8.64}{\theta}\left(\frac{2\theta^3}{8} - \frac{2\theta^3}{27}\right) = \frac{38\theta^2}{25},$$

所以方差为

$$\frac{38\theta^2}{25} - \theta^2 = \frac{13\theta^2}{25} > \frac{\theta^2}{3}.$$

样本均值有更小的 MSE 而且这个结果与 θ 的真实值无关, 所以在这个问题中它是 θ 的优良估计. □

例 9.15 对于区间 $(0,\theta)$ 上的均匀分布, 比较估计量 $2\bar{X}$ 和 $[(n+1)/n]\max(X_1,\cdots,X_n)$ 的 MSE. 同时计算 $\max(X_1,\cdots,X_n)$ 的 MSE.

解 前两个估计变量都是无偏的, 所以只要比较它们的方差就足够了. 对于样本均值的两倍, 有

$$\mathrm{Var}(2\bar{X}) = \frac{4}{n}\mathrm{Var}(X) = \frac{4\theta^2}{12n} = \frac{\theta^2}{3n}.$$

调整后最大值的二阶矩为

$$\mathrm{E}\left[\left(\frac{n+1}{n}Y_n\right)^2\right] = \frac{(n+1)^2}{n^2}\frac{n\theta^2}{n+2} = \frac{(n+1)^2\theta^2}{(n+2)n},$$

方差为

$$\frac{(n+1)^2\theta^2}{(n+2)n} - \theta^2 = \frac{\theta^2}{n(n+2)}.$$

除了 $n=1$ 的情况 (此时两个估计量等价), 基于最大观测的估计量有较小的 MSE. 第三个估计量是有偏的, 它的 MSE 为

$$\frac{n\theta^2}{(n+2)(n+1)^2} + \left(\frac{n\theta}{n+1} - \theta\right)^2 = \frac{2\theta^2}{(n+1)(n+2)},$$

也是比调整后最大观测值的 MSE 要大. □

习题

9.1 在例 9.1 中, 无放回的抽取 3 个样本, 证明样本均值是总体均值的无偏估计, 而样本中位数是总体均值的有偏估计.

9.2 证明: 对于随机抽取的样本, 样本均值总是总体均值的无偏估计.

9.3 设 X 服从区间 $(\theta-2,\theta+2)$ 上的均匀分布, 也就是说 $f_X(x)=0.25, \theta-2<x<\theta+2$. 证明样本量为 3 时, 样本中位数是 θ 的无偏估计.

9.4 对于 Pareto 分布, 解释为什么样本均值可能并不是总体均值的相合估计.

9.5 习题 9.3 中样本量为 3, 比较样本均值和样本中位数作为 θ 的估计时的 MSE.

9.6* 给定未知参数 θ 的两个独立的估计量. 对于估计量 A, 有: $\mathrm{E}(\hat{\theta}_A)=1\ 000$, $\mathrm{Var}(\hat{\theta}_A)=160\ 000$. 对于估计量 B, 有: $\mathrm{E}(\hat{\theta}_B)=1\ 200$,$\mathrm{Var}(\hat{\theta}_B)=40\ 000$. 估计量 C 是 A 和 B 的加权平均, $\hat{\theta}_C=\omega\hat{\theta}_A+(1-\omega)\hat{\theta}_B$, 计算 ω 的值使 $\mathrm{Var}(\hat{\theta}_C)$ 最小.

9.7* 损失总量服从 Pareto 分布 (见附录 A), $\theta=6\ 000$, α 未知. 基于 10 个样本的最大似然估计的模拟结果显示 $\mathrm{E}(\hat{\alpha})=2.2$, $\mathrm{MSE}(\hat{\alpha})=1$, 若已知 $\alpha=2$, 计算 $\mathrm{Var}(\hat{\alpha})$.

9.8* 现使用 2 种工具测量非零距离, 随机变量 X 表示第一种工具的测量值, Y 表示第二种工具的测量值. 假设 X 和 Y 独立, $\mathrm{E}(X)=0.8m$, $\mathrm{E}(Y)=m$, $\mathrm{Var}(X)=m^2$, $\mathrm{Var}(Y)=1.5m^2$, 其中 m 是真实距离. 考虑 m 的估计量 $Z=\alpha X+\beta Y$, 计算 α 和 β 的值, 使 Z 为这类估计量中的 UMVUE.

9.9 已知总体有 6 个值, 分别为: 1, 1, 2, 3, 5, 10. 无放回的随机抽取 3 个样本, 在以下每种情况下分别估计总体均值. 注: 利用电子表和最优化方法可以很好地解决这个问题.

(a) 计算样本均值的偏差、方差和 MSE.

(b) 计算样本中位数的偏差、方差和 MSE.

(c) 计算样本中列数 (最大观测和最小观测的平均) 的偏差、方差和 MSE.

(d) 考虑任何一个形如 $aX_{(1)}+bX_{(2)}+cX_{(3)}$ 的估计量, 其中 $X_{(1)}\leqslant X_{(2)}\leqslant X_{(3)}$ 为样本的次序统计量.

i. 给出 a,b,c 的一个约束条件, 使得估计量是无偏的.

ii. 确定 a,b,c 的值, 得到一个最小方差无偏估计量.

iii. 确定 a,b,c 的值, 得到一个 MSE 最小的估计量 (有可能是有偏的).

9.10* 考虑 2 个不同的估计量 $\hat{\theta}_1$ 和 $\hat{\theta}_2$, 为了对它们进行检测, 在真实值为 $\theta=2$ 的情况下进行了 75 次模拟试验, 结果如下:

$$\sum_{j=1}^{75}\hat{\theta}_{1j}=165,\quad \sum_{j=1}^{75}\hat{\theta}_{1j}^2=375,\quad \sum_{j=1}^{75}\hat{\theta}_{2j}=147,\quad \sum_{j=1}^{75}\hat{\theta}_{2j}^2=312,$$

其中的 $\hat{\theta}_{ij}$ 为估计量 $\hat{\theta}_i$ 在第 j 次试验中得到的估计值. 对每个估计量计算 MSE, 并计算**相对有效性**(MSE 的比率).

9.3 区间估计

到目前为止所讨论的估计量均为点估计, 也就是说, 通过估计过程产生一个数值, 它表示我们对未知总体的最好尝试结果. 即便这个值估计的很好, 也无法期望它完全匹配真实值. 而更加有效的陈述通常是通过**区间估计**得到的. 估计过程的结

果为一种取值的范围, 其中的任何一个值都有可能是真实值, 而不是一个唯一单独的数值. 区间估计的一个具体形式是置信区间.

定义 9.16 参数 θ 的 $100(1-\alpha)\%$ 置信区间为一对随机取值 L 和 U, 它们在随机样本的计算中满足: 对任意的 θ, 有 $\Pr(L \leqslant \theta \leqslant U) \geqslant 1-\alpha$.

注意这个定义并没有给出一个唯一的区间. 因为这个定义是一种概率描述, 必须对所有 θ 成立, 它没有说明一个具体区间是否包含来自总体的 θ 的真实值. 置信水平 $1-\alpha$ 反映了获得 U 和 L 的方法的置信水平, 而不是表示具体值. 恰当的解释是: 如果用这个区间估计的方法不断对各种样本进行估计, 所得到的区间至少在 $100(1-\alpha)\%$ 的区间内包括真实值.

建立置信区间通常非常困难. 例如我们知道, 如果总体服从均值和方差未知的正态分布, 均值的 $100(1-\alpha)\%$ 置信区间为

$$L = \bar{X} - t_{\alpha/2,n-1}s/\sqrt{n}, \quad U = \bar{X} + t_{\alpha/2,n-1}s/\sqrt{n}, \tag{9.2}$$

其中$s = \sqrt{\sum\limits_{j=1}^{n}(X_j - \bar{X})^2/(n-1)}$, $t_{\alpha/2,b}$ 是 t 分布 (b 个自由度) 的 $100(1-\alpha/2)\%$ 分位点. 但需要很多的工作来证明这个结论 (可见 [58] 的第 214 页相关内容).

然而, 也可以用另一种方法来构造近似的置信区间. 假设已知参数 θ 的点估计 $\hat{\theta}$, 满足 $\mathrm{E}(\hat{\theta}) \doteq \theta$, $\mathrm{Var}(\hat{\theta}) \doteq v(\theta)$. $\hat{\theta}$ 近似服从正态分布, 定理 12.13 将证明这是常见的情况. 根据上述的这些近似, 有

$$1-\alpha \doteq \Pr\left(-z_{\alpha/2} \leqslant \frac{\hat{\theta}-\theta}{\sqrt{v(\theta)}} \leqslant z_{\alpha/2}\right), \tag{9.3}$$

其中 $z_{\alpha/2}$ 是标准正态分布的 $100(1-\alpha/2)\%$ 分位点, 对上式求解 θ 就可得到所需的区间. 有时这样做是很困难的 (由于 θ 出现在分母中), 所以如果必要可以将 (9.3) 式中的 $v(\theta)$ 改为 $v(\hat{\theta})$, 从而得到如下近似,

$$1-\alpha \doteq \Pr\left(\hat{\theta} - z_{\alpha/2}\sqrt{v(\hat{\theta})} \leqslant \theta \leqslant \hat{\theta} + z_{\alpha/2}\sqrt{v(\hat{\theta})}\right). \tag{9.4}$$

例 9.17 对于方差未知的正态分布, 利用 (9.4) 式构造均值的 95%近似置信区间.

解 利用 $\hat{\theta} = \bar{X}$, 注意 $\mathrm{E}(\hat{\theta}) = \theta$, $\mathrm{Var}(\hat{\theta}) = \sigma^2/n$, 并且 $\hat{\theta}$ 服从正态分布, 所以置信区间为 $\bar{X} \pm 1.96s/\sqrt{n}$. 因为 $t_{0.025,n-1} > 1.96$, 这个区间一定比 (9.2) 式给出的准确区间窄, 这意味着我们的置信水平有时小于 95%. □

例 9.18 对于 Poisson 分布, 在 $n = 25, \bar{x} = 0.12$ 的情况下, 分别利用 (9.3) 式和 (9.4) 式构造均值的 95%近似置信区间.

解　令 $\hat{\theta}=\bar{X}$ 为样本均值, Poisson 分布满足 $\mathrm{E}(\hat{\theta})=\mathrm{E}(X)=\theta$ 和 $v(\theta)=\mathrm{Var}(\bar{X})=\mathrm{Var}(X)/n=\theta/n$, 所以第一个区间

$$0.95 \doteq \Pr\left(-1.96 \leqslant \frac{\bar{X}-\theta}{\sqrt{\theta/n}} \leqslant 1.96\right)$$

成立当且仅当

$$|\bar{X}-\theta| \leqslant 1.96\sqrt{\frac{\theta}{n}},$$

其等价于

$$(\bar{X}-\theta)^2 \leqslant \frac{3.841\,6\theta}{n}$$

或

$$\theta^2-\theta\left(2\bar{X}+\frac{3.841\,6}{n}\right)+\bar{X}^2 \leqslant 0.$$

求解二次方程, 得到区间

$$\bar{X}+\frac{1.920\,8}{n} \pm \frac{1}{2}\sqrt{\frac{15.366\,4\bar{X}+3.841\,6^2/n}{n}}.$$

因此本题的区间为 0.197±0.156.

第二个近似区间为 $\bar{X} \pm 1.96\sqrt{\bar{X}/n}$, 具体本题为 0.12±0.136. 这个区间包含了小于零的部分 (对于 θ 的真实值这是不可能的), 这是因为本题采用 (9.4) 式进行近似过于粗糙. □

习题

9.11　$x_1,\cdots,x_n$ 是随机样本, 概率密度函数为 $f(x)=\theta^{-1}\mathrm{e}^{-x/\theta}$, $x>0$. 这个指数分布的均值为 θ, 方差为 θ^2. 考虑样本均值 $\bar{X}$ 作为 θ 的估计量, 可得出 $\bar{X}/\theta$ 服从 $\alpha=n,\theta=1/n$ 的 gamma 分布, 其中第二个表达式中左边的 θ 是 gamma 分布的参数. 对于样本量为 50, 样本均值为 275 的情况, 按照下面的方法分别建立 95%置信区间. 如果方程中需要 θ 的真实值, 则用估计值代替.

(a) 利用 gamma 分布确定准确的区间.

(b) 使用正态近似, 在求解 (9.4) 式的不等式之前估计方差.

(c) 使用正态近似, 在求解例 9.18 中的不等式之后估计 θ.

9.4　假 设 检 验

在大多数数理统计教科书中都会详细介绍假设检验, 所以这部分的介绍相当直接, 不涉及统计假设检验基本原理中的哲学思想, 也不考虑其他方法. 假设检验问题首先要有 2 个假设: 一个称为零假设; 另一个称为备择假设. 按照传统的符号,

H_0 表示零假设, H_1 表示备择假设. 这两个假设不是对称的, 对换它们的位置可能会改变推断结果. 下面用一个简单的例子说明这个过程.

例 9.19 你所在的保险公司原来基于平均索赔为 1 200 的假设制定保费价格. 你希望提高保费价格. 监管者要求公司提供证据, 证明公司目前的平均索赔超过 1 200. 为了提供这种证据, 你收集了以下数据. 针对这个问题你应该设计什么样的检验假设?

27	82	115	126	155	161	243	294	340	384
457	680	855	877	974	1 193	1 340	1 884	2 558	15 743

解 令 μ 表示总体均值. 一种假设 (你声称它是正确的) 为 $\mu > 1\ 200$. 因为一个假设检验必须提供一个二者择一的情况, 所以另一个假设必然为 $\mu \leqslant 1\ 200$, 剩下的问题就是决定哪一个是零假设. 当连续分布的总体被划分为 2 部分时, 就可能存在边界需要被指定到其中一个假设中. 而包含边界的假设必须是零假设, 所以这个问题可以简单地表述为:

$$H_0: \mu \leqslant 1\ 200,$$

$$H_1: \mu > 1\ 200. \qquad \square$$

假设检验的结论是通过计算称为**检验统计量**的值来确定的. 检验统计量是观测值的函数, 可看作随机变量. 也就是说, 我们在设计检验程序时, 关心的是可能得到的样本, 而不是已经得到的具体的样本. 检验规则是通过建立否定域实现的, 它是检验统计量可能取值的一个子集. 如果观测到的样本使检验统计量的值落在否定域中, 就否定零假设, 而备择假设就是数据所支持的结论, 否则零假设不能被否定 (后面会有更多叙述). 否定域的边界 (除了 $\pm\infty$) 称为临界值.

例 9.20(续例 9.19) 使用大部分统计书中介绍的检验统计量和否定域进行检验. 假设总体服从标准差为 3 435 的正态分布.

解 本题的传统检验统计量为

$$z = \frac{\bar{x} - 1\ 200}{3\ 435/\sqrt{20}} = 0.292,$$

如果 $z > 1.645$ 则拒绝零假设. 因为 0.292 小于 1.645, 所以零假设不能被拒绝, 数据不支持平均索赔超过 1 200 的看法. $\square$

前面例子中构造的检验需要满足某些目标. 第一个目标是控制所谓的第一类错误, 它是零假设正确的情况下拒绝零假设的错误. 在这个例子中, 很多情况零假设都是正确的, 从而引出一个常用的度量标准来检验犯第一类错误的倾向.

定义 9.21 假设检验的显著性水平是已知零假设正确的条件下犯第一类错误的概率. 如果零假设在不止一种情况下都正确, 则显著性水平是这些概率的最大值. 显

著性水平通常用字母 α 表示.

这是一个考虑最坏情况的保守性定义, 典型的情况就是 2 个假设间存在边界.

例 9.22 计算例 9.20 中的显著性水平.

解 首先计算 $\mu = 1\ 200$ 零假设正确时犯第一类错误的概率, 即

$$\Pr(Z > 1.645|\mu = 1\ 200) = 0.05.$$

这是因为假设 Z 服从标准正态分布.

现在假设 μ 的值小于 1 200, 则

$$\begin{aligned}&\Pr\left(\frac{\bar{X} - 1\ 200}{3\ 435/\sqrt{20}} > 1.645\right) = \Pr\left(\frac{\bar{X} - \mu + \mu - 1\ 200}{3\ 435/\sqrt{20}} > 1.645\right)\\ &= \Pr\left(\frac{\bar{X} - \mu}{3\ 435/\sqrt{20}} > 1.645 - \frac{\mu - 1\ 200}{3\ 435/\sqrt{20}}\right).\end{aligned}$$

因为已知 μ 小于 1 200, 所以右边总是大于 1.645, 而左边服从标准正态分布, 所以概率小于 0.05, 即显著性水平为 0.05. □

显著性水平通常是预先给定的, 一般在 1%到 10%之间. 第二个目标就是要使犯第二类错误 (在备择假设正确时没有拒绝零假设) 的概率尽量小. 通常试图减少犯某一类错误的概率的同时总会增加犯另一类错误的概率, 所以一旦确定了显著性水平, 我们最多能做的就是使犯第二类错误的概率尽可能的小, 尽管并不能保证这是一个很小的数. 最佳的检验要满足如下要求.

定义 9.23 *称假设检验为一致最大功效的, 如果不存在其他检验, 使得显著性水平小于或等于它, 且对备择假设范围中的任意一个值, 相比于其他检验使用这个检验有更小的犯第二类错误的概率.*

例 9.24(续例 9.22) 当 $\mu = 2\ 000$ 备择假设正确时, 计算犯第二类错误的概率.

解

$$\begin{aligned}&\Pr\left(\frac{\bar{X} - 1\ 200}{3\ 435/\sqrt{20}} < 1.645|\mu = 2\ 000\right)\\ &= \Pr(\bar{X} - 1\ 200 < 1\ 263.51|\mu = 2\ 000)\\ &= \Pr(\bar{X} < 2\ 463.51|\mu = 2\ 000)\\ &= \Pr\left(\frac{\bar{X} - 2\ 000}{3\ 435/\sqrt{20}} < \frac{2\ 463.51 - 2\ 000}{3\ 435/\sqrt{20}} = 0.603\ 5\right) = 0.726\ 9.\end{aligned}$$

对于这个 μ 值, 检验不是很有功效, 因为有超过 70%的机会犯第二类错误. 然而 (尽管不容易证明), 对这个问题来说, 这里的检验是一致最大功效的. □

因为犯第二类错误的概率可能很高, 所以一般来说, 当零假设没有被拒绝时无法给出很明确的结论. 这时, 与 "选择或接受零假设" 的说法相比较 "不能拒绝零

假设" 的说法更合适. 也就是说, 我们没有从样本中得到充分的证据支持备择假设, 所以得不到任何实质性结论.

对这种假设检验方法存在的一种普遍的批评意见是显著性水平的选择是任意的. 事实上, 通过改变显著性水平, 可以得到任何需要的结果.

例 9.25(续例 9.24) 使用显著性水平 $\alpha = 0.45$ 完成检验, 然后分别计算零假设被拒绝和不被拒绝时显著性水平的范围.

解 因为 $\Pr(Z > 0.125\ 7) = 0.45$, 所以当满足

$$\frac{\bar{X} - 1\ 200}{3\ 435/\sqrt{20}} > 0.125\ 7.$$

时, 拒绝零假设. 在这个例子中, 检验统计量为 0.292, 落在拒绝域中, 所以零假设被拒绝. 当然, 很少有人会用 45%作为检验结果的置信度. 因为 $\Pr(Z > 0.292) = 0.385\ 1$, 所以如果选择显著性水平大于 38.51%则拒绝零假设, 如果选择的显著性水平小于 38.51%则不拒绝零假设. □

很少有人愿意在 38.51%的时间都出错. 基于这个显著水平的结论应该比前面基于 5%显著性水平的结论更具有说服力. 当采用某个显著性水平进行检验时, 读者应该同时考虑在其他显著性水平下会得到什么结果. 这里的 38.51%称为 p 值, 其定义如下.

定义 9.26 *在假设检验中, p 值是检验统计量比从样本中得到的值更不符合零假设的概率. 如果检验前确定的显著性水平大于 p 值, 则拒绝零假设; 如果检验前确定的显著性水平小于 p 值, 则不能拒绝零假设.*

因为 p 值必须在 0 和 1 之间, 它的数值大小也具有一定的含义. 这个值越靠近零, 说明数据越支持备择假设. 在一般的实践中, 如果这个值大于 10%, 则意味着数据不能提供证据来支持备择假设; 如果 p 值小于 1%, 则说明强烈支持备择假设. 如果值在两者之间, 则说明还不能肯定得到一个合适的结论, 可能需要更多的数据, 也可能需要对数据作更仔细的分析, 或者需要进一步的实验来得到结论.

习题

9.12 (续习题 9.11) 检验 $H_0: \theta \geqslant 325$ 与 $H_1: \theta < 325$, 显著水平 5%, 检验统计量为样本均值. 并且分别利用检验统计量的精确分布和正态近似计算 p 值.

第10章　基于完整数据的统计估计

10.1　引　　言

本章及第 11 章的内容, 在以往的精算教材中是列在所谓的 “生存模型” 之中的, 这使人们感觉这些技术只是在研究生存时间的概率分布时才会使用. Klein and Moeschberger[74] 以及 Lawless[81] 所著的关于这个主题的标准教科书中的实际例题也完全倾向于此. 不过, 正如在第 10 章和第 11 章中我们将会看到的, 在生存时间建模过程中遇到的许多问题也同样存在于一般赔付的建模过程中. 我们给出的例题既有生存时间模型方面的, 也有一般赔付模型方面的. 为了强调这一点, 带星号的习题取材于以往的 SOA 课程 160 的考试真题, 但题目的背景已经变为与一般赔付额有关的问题. 书后只列出了关于这部分内容的少数参考文献, 大部分结论都在生存模型的文献中得到了充分的论述. 希望了解更多细节和证明过程的读者应当去查阅专门论述生存模型的教材, 比如上文中提及的教材.

本章假设生存模型的类型已知, 但模型的其他一些特征可能是未知的. 在第 4 章中, 生存模型被分为两类 —— 数据依赖的模型 (data-dependent distribution) 和参数模型 (parametric distribution). 这里再重复一下这两类模型的定义:

定义 10.1　*数据依赖型分布是指这样一类分布, 它的复杂程度至少与产生它的数据或者其他信息集相当, 并且其 “参数” 的个数会随着数据点或者信息量的增加而增加.*

定义 10.2　*参数型分布是指某个分布函数族, 其中的每个分布都由一个或者多个数值唯一确定, 称该数值为 “参数”. 参数的个数必须是有限常数.*

这里只考虑两种数据依赖型分布, 它们对数据的依赖方式是相似的. 以下是这两种模型的最简洁的定义.

定义 10.3　*经验分布* (empirical distribution) *设每个数据点的概率为 $1/n$, 即得到经验分布.*

定义 10.4　*核光滑分布* (kernel smoothed distribution) *将每个数据点看作某个连续型随机变量的取值, 并令取到每个随机变量的概率为 $1/n$, 这样的分布称为核光滑分布. 每个数据点对应的随机变量分布除位置参数和形状参数外是相同的.*

值得注意的是, 经验分布是核光滑分布的一个特例: 只要将核光滑分布的每个随机变量的分布取为在该数据点的退化分布, 就得到了经验分布. 以后还会遇到和

经验分布本质相同但产生不同数值的分布. 第 11 章的内容将告诉大家为了考虑删失数据 (censored data) 和截断数据 (truncated data), 应当如何修正经验分布的定义. 关于核光滑处理中可选的分布函数, 11.3 节介绍了其中的几种.

本章的讨论反复用到了以下 4 个例子. 由于它们代表了精算实务中的典型数据, 下文中一律将其简称为数据集 A、数据集 B、数据集 C 和数据集 D.

数据集 A　这个数据集经常出现在非寿险精算的文献中. 它首次出现于 1959 年 Dropkin 的论文 [30] 中. Dropkin 收集了 1956—1958 年间 94 935 个驾驶员每人每年出现交通事故数的数据, 如表 10-1 所示.

表 10-1　数据集 A

事故次数	驾驶员人数
0	81 714
1	11 306
2	1 618
3	250
4	40
5 次以上	7

数据集 B　这是一组人造的工伤险赔付数据, 与任何具体的保单或被保险群体无关. 而且赔付都是按照损失量全额支付的. 表 10-2 列出了 20 个随机生成的赔付额数据样本.

表 10-2　数据集 B

27	82	115	126	155	161	243	294	340	384
457	680	855	877	974	1 193	1 340	1 884	2 558	15 743

数据集 C　观测值是一组普通责任保险保单的 227 例赔案的赔付额, 列于表 10-3.

表 10-3　数据集 C

赔付额范围	赔付笔数
0~7 500	99
7 500~17 500	42
17 500~32 500	29
32 500~67 500	28
67 500~125 000	17
125 000~300 000	9
300 000 以上	3

数据集 D　这是一组人造的 5 年定期寿险保单中止时间的数据集. 保单的中止或

者是因被保险人身故, 或者是因被保人退保 (保单合约的解除), 或者是 5 年保单满期. 因此数据分两个表给出. 数据集 D1(表 10-4) 列出了从发行之日起观测的 30 份保单的数据, 不仅详细给出了每个投保人的身故时间还给出了退保时间 (只要身故或退保发生在 5 年到期之前). 当然, 通常情况下我们无法得知已退保的投保人的确切身故时间, 也无法得知某个事实上已经身故的投保人如果没有身故将在何时退保. 而表中最后的 12 个投保人在 5 年内既未身故也没有退保, 保单一直到满期.

表 10-4　数据集 D1

保单持有人编号	身故时间	退保时间
1	—	0.1
2	4.8	0.5
3	—	0.8
4	0.8	3.9
5	3.1	1.8
6	—	1.8
7	—	2.1
8	—	2.5
9	—	2.8
10	2.9	4.6
11	2.9	4.6
12	—	3.9
13	4.0	—
14	—	4.0
15	—	4.1
16	4.8	—
17	—	4.8
18	—	4.8
19～30	—	—

数据集 D2(表 10-5) 对前 30 个投保人记录了某个事件 (或死亡、或退保) 的首次发生时间, 最后新增的 10 位保单持有人的观测是从保单发行一段时间后才开始的. 表 10-5 汇总了所有 40 张保单的情况. “初始观测时间” 一栏给出了每张保单的开始观测时间; “最后观测时间” 一栏给出了每个保单最后一次观测的时间 (以上 2 个时间均以保单发行为起点); “事件” 一栏中 “s” 表示退保, “d” 表示身故, “e” 表示 5 年后保单满期.

如果能够对概率分布的任意点收集数据, 则理想的状况是得到每个观测值 (本质上) 的精确值①, 这种情况称作 “完整个体数据”(complete individual data), 数据

① 有些数值是不可能精确度量的, 比如年龄只能四舍五入到整数, 货币金额四舍五入到元, 车辆行驶里程四舍五入到英里等等. 本书并不关心这些取整造成的误差, 而是将处理后的近似值作为精确值看待.

集 B 和数据集 D1 就属于这种情况. 一般来说无法得到精确值主要有 2 个原因. 一个原因是分组处理 —— 人们作记录时往往只知道观测值属于某个范围而不是具体的值. 数据集 C 以及数据集 A 中 "5 次以上" 正是指这种情况.

表 10-5 数据集 D2

保单组合	初始观测时间	最后观测时间	事件	保单组合	初始观测时间	最后观测时间	事件
1	0	0.1	s	16	0	4.8	d
2	0	0.5	s	17	0	4.8	s
3	0	0.8	s	18	0	4.8	s
4	0	0.8	d	19~30	0	5.0	e
5	0	1.8	s	31	0.3	5.0	e
6	0	1.8	s	32	0.7	5.0	e
7	0	2.1	s	33	1.0	4.1	d
8	0	2.5	s	34	1.8	3.1	d
9	0	2.8	s	35	2.1	3.9	s
10	0	2.9	d	36	2.9	5.0	e
11	0	2.9	d	37	2.9	4.8	s
12	0	3.9	s	38	3.2	4.0	d
13	0	4.0	d	39	3.4	5.0	e
14	0	4.0	s	40	3.9	5.0	e
15	0	4.1	s				

另一个无法获得精确值的原因是可能有删失或者截断发生. 左删失数据是指只知道观测值在某个给定值之下而不知道其具体值; 右删失数据是指只知道观测值在某个给定值之上同样也不知道其具体取值. 实际上删失数据已经对数据进行了分组, 如果数据已经分组, 再考虑删失就没有意义了. 例如, 在数据集 C 中, 也许在 300 000 点右删失, 但是我们无法从数据本身确认这一点, 这方面的信息也不会影响我们处理这批数据的方法. 另外, 如果数据 B 在 1 000 点右删失, 我们只能得到 15 个具体的观测值, 另外 5 个位于 1 000 到无穷大区间上的观测值则合并为一个组.

在保单设计中经常出现右删失的情况. 例如, 若保单对任何事故的赔付额都不超过 100 000 元, 那么赔付额数据本身将无从考查那些事故的损失量超过 100 000 的保单的实际损失量, 不过我们可以确认这样的事故确实发生过. 数据集 D2 是随机删失的. 考虑表 10-5 中的第 5 个保单: 如果无法获得 "其他信息", 那么关于身故时间的信息我们唯一能获知的是身故发生在保单发行 1.8 年之后. 所有保单的数据都在第 5 年右删失, 这是由保单本身的属性决定的. 同时注意到数据集 A 也是

在 5 这点右删失的. 这种叙述比 "数据集 A 既有个体数据也有分组数据" 的说法更常用.

左截断数据是低于某个给定值的观测值不做记录; 而右截断意味着对高于某个给定值的观测值不做记录. 在保单设计时经常会考虑左截断的处理. 对于 250 元免赔额的车险保单, 保险公司不会过问损失额低于 250 元的汽车损伤情况, 也不会记录这些损伤的具体数据. 数据集 D2 中编号为 31~40 的观测值就是在不同的数值点做左截断处理. 对任何数据集合都可以强制进行截断处理, 比如, 若要对数据集 B 按 250 处进行左截断处理, 只需将前面 7 个观测值去掉并保持剩余的 13 个观测值不变即可.

10.2　完整个体数据的经验分布

由定义 10.3 我们看到, 在经验分布中每个数据点的概率都是 $1/n$. 当每个数据点都有记录时, 这样做是有效的. 以下是经验分布的另一个等价定义.

定义 10.5　*经验分布函数为*

$$F_n(x) = \frac{\text{观测值} \leqslant x \text{ 的个数}}{n},$$

其中 n 是总观测数.

例 10.6　*求数据集 A 和数据集 B 的经验概率函数以及数据集 A 的经验分布函数. 假设数据集 A 中 7 位事故次数为 "5 次或 5 次以上" 的驾驶员, 其事故次数恰为 5 次.*

解　为了使记号简便, 用脚标中的样本总数 (如果样本总数未知总用 n 表示) 表示相应数据集的经验函数. 如果不写脚标, 则表示随机变量的真实概率函数或分布函数. 对于数据集 A, 概率函数估计为

$$p_{94\,935}(x) = \begin{cases} 81\,714/94\,935 = 0.860\,736, & x = 0, \\ 11\,306/94\,935 = 0.119\,092, & x = 1, \\ 1\,618/94\,935 = 0.017\,043, & x = 2, \\ 250/94\,935 = 0.002\,633, & x = 3, \\ 40/94\,935 = 0.000\,421, & x = 4, \\ 7/94\,935 = 0.000\,074, & x = 5, \end{cases}$$

其中各个四舍五入后的概率值之和为 0.999 999. 其分布函数是在每个数据点跳跃

的阶梯函数.

$$F_{94\ 935}9(x)=\begin{cases}0/94\ 935=0.000\ 000, & x<0,\\ 81\ 714/94\ 935=0.860\ 736, & 0\leqslant x<1,\\ 93\ 020/94\ 935=0.979\ 828, & 1\leqslant x<2,\\ 94\ 638/94\ 935=0.996\ 872, & 2\leqslant x<3,\\ 94\ 888/94\ 935=0.999\ 505, & 3\leqslant x<4,\\ 94\ 928/94\ 935=0.999\ 926, & 4\leqslant x<5,\\ 94\ 935/94\ 935=1.000\ 000, & x\geqslant 5.\end{cases}$$

对于数据集 B, 有

$$p_{20}(x)=\begin{cases}0.05, & x=27,\\ 0.05, & x=82,\\ 0.05, & x=115,\\ \vdots & \vdots\\ 0.05, & x=15\ 743.\end{cases}$$

□

正如在这个例子中所看到的, 经验模型是离散分布模型. 因此, 根据这个模型, 我们无法通过求导来得到密度函数和风险率函数 (hazard rate function). 我们能够得到与风险率函数最接近的指标是下面定义的累积风险率函数 (cumulative hazard rate function).

定义 10.7 *累积风险率函数定义为*

$$H(x)=-\ln S(x).$$

之所以将其取名为 "累积风险率函数", 是因为如果 $S(x)$ 可导, 则有

$$H'(x)=-\frac{S'(x)}{S(x)}=\frac{f(x)}{S(x)}=h(x),$$
$$H(x)=\int_{-\infty}^{x}h(y)\mathrm{d}y.$$

分布函数可由 $F(x)=1-S(x)=1-\mathrm{e}^{-H(x)}$ 得到. 进而, 对分布函数的估计有时也可通过估计累积风险率函数来实现.

为了定义经验估计, 还需要引入一些记号. 对一个容量为 n 的样本, 设 $y_1<y_2<\cdots<y_k$ 为样本中出现的所有不同的观测值, 这里 k 一定要小于或者等于 n. 记观测值 y_j 在样本中出现的次数为 s_j. 此时必有 $\sum_{j=1}^{k}s_j=n$. 另一个关心的量是数据集合中不小于某一个给定数值的观测值的个数. 这样的观测值及其个数都称

为**风险集**(risk set). 用 $r_j=\sum\limits_{i=j}^{k}s_i$ 表示不小于 y_j 的观测值的个数. 用这个记号可将经验分布函数表示为

$$F_n(x)=\begin{cases}0, & x<y_1,\\ 1-\frac{r_j}{n}, & y_{j-1}\leqslant x<y_j,\quad j=2,\cdots,k,\\ 1, & x\geqslant y_k.\end{cases}$$

例 10.8 考虑由数值 1.0, 1.3, 1.5, 1.5, 2.1, 2.1, 2.1, 2.8 构成的数据. 计算上一个自然段中定义的各个量, 并写出经验分布函数.

解 由于样本中共有 5 个不同的值, 所以 $k=5$. y_j, s_j 以及 r_j 的值如表 10-6 所示.

表 10-6 例 10.8 中的值

j	y_j	s_j	r_j
1	1.0	1	8
2	1.3	1	7
3	1.5	2	6
4	2.1	3	4
5	2.8	1	1

$$F_8(x)=\begin{cases}0, & x<1.0\\ 1-\frac{7}{8}=0.125, & 1.0\leqslant x<1.3,\\ 1-\frac{6}{8}=0.250, & 1.3\leqslant x<1.5,\\ 1-\frac{4}{8}=0.500, & 1.5\leqslant x<2.1,\\ 1-\frac{1}{8}=0.875, & 2.1\leqslant x<2.8,\\ 1, & x\geqslant 2.8.\end{cases}$$ □

定义 10.9 累积风险率函数的 Nelson-Åalen 估计 ([1], [99]) 定义为

$$\hat{H}(x)=\begin{cases}0, & x<y_1,\\ \sum\limits_{i=1}^{j-1}\frac{s_i}{r_i}, & y_{j-1}\leqslant x<y_j, j=2,\cdots,k,\\ \sum\limits_{i=1}^{k}\frac{s_i}{r_i}, & x\geqslant y_k.\end{cases}$$

由于这是一个阶梯函数, 它的导数 (可以给出风险率函数的估计) 并没有特别的意义. 如果读者希望了解该估计方法原创者的直观想法, 请参阅本书第 11 章例 11.3 的相关内容.

例 10.10 计算例 10.8 中数据的 Nelson-Åalen 估计.

解 经计算得

$$\hat{H}(x)=\begin{cases}0, & x<1.0,\\ \frac{1}{8}=0.125, & 1.0\leqslant x<1.3,\\ 0.125+\frac{1}{7}=0.268, & 1.3\leqslant x<1.5,\\ 0.268+\frac{2}{6}=0.601, & 1.5\leqslant x<2.1\\ 0.601+\frac{3}{4}=1.351, & 2.1\leqslant x<2.8,\\ 1.351+\frac{1}{1}=2.351, & x\leqslant 2.8.\end{cases}$$

运用这些数值, 通过指数运算可得到分布函数的估计. 例如, 对 $1.5\leqslant x<2.1$, 分布函数的估计值是 $\hat{F}(x)=1-\mathrm{e}^{-0.601}$=0.452, 并不等于经验估计值 0.5. 下文中都用函数符号上加一个帽子 (∧) 表示不同于经验分布方法估计的函数值. □

例 10.11 **根据数据集 D1 提供的数据计算经验生存函数以及身故时间的累积风险率函数的 Nelson-Åalen 估计. 用 Nelson-Åalen 方法估计生存函数. 假设退保后的身故时间是已知的.**

解 计算结果如表 10-7 所示, 其中经验函数对应的各个值是基于从当前的 y 值到下一个 y 值的左闭右开的区间计算的. □

表 10-7 例 10.11 的数据

j	y_j	s_j	r_j	$S_{30}(x)$	$\hat{H}(x)$	$\hat{S}(x)=\mathrm{e}^{-\hat{H}(x)}$
1	0.8	1	30	$\frac{29}{30}=0.966\ 7$	$\frac{1}{30}=0.033\ 3$	0.967 2
2	2.9	2	2.9	$\frac{27}{30}$=0.900 0	0.033 3+$\frac{2}{29}$=0.102 3	0.902 8
3	3.1	1	27	$\frac{26}{30}$=0.866 7	0.102 3+$\frac{1}{27}$=0.139 3	0.870 0
4	4.0	1	26	$\frac{25}{30}$=0.833 3	0.139 3+$\frac{1}{26}$ = 0.177 8	0.837 1
5	4.8	2	25	$\frac{23}{30}$=0.766 7	0.177 8+$\frac{2}{25}$=0.257 8	0.772 7

在这个具体问题中, 由于已知所有的保单都在 5 年内中止, 因此 5 年以后的结果没有实际意义. 以上介绍的经验分布方法仅仅适用于能够得到个别观测数据且没有删失和截断的情形, 第 11 章将介绍在有删失和截断的情形下该模型应当如何调整.

习题

10.1 对数据集 D1, 求退保时间的经验分布函数并运用 Nelson- Åalen 方法估计退保时间的分布函数. 假设已身故的投保人的退保时间是已知的.

10.2 表 10-8 中的数据来自 Loss Distribution[59] 的第 128 页. 这些数据记录了 1949 年到 1980 年之间的 35 次飓风造成的趋势调整总损失额, 所有损失额均已经通货膨胀调整 [用

6
文

住房指数 (Residential Construction Index, RCI) 调整] 为与 1981 年美元币值等价的金额. 表中记载了所有趋势调整的飓风损失额在 5 000 000 以上的数据.

表 10-8 飓风造成的趋势损失额

发生年份	损失额 (10^3)	发生年份	损失额 (10^3)	发生年份	损失额 (10^3)
1964	6 766	1964	40 596	1975	192 013
1968	7 123	1949	41 409	1972	198 446
1971	10 562	1959	47 905	1964	227 338
1956	14 474	1950	49 397	1960	329 511
1961	15 351	1954	52 600	1961	361 200
1966	16 983	1973	59 917	1969	421 680
1955	18 383	1980	63 123	1954	513 586
1958	19 030	1964	77 809	1954	545 778
1974	25 304	1955	102 942	1970	750 389
1959	29 112	1967	103 217	1979	863 881
1971	30 146	1957	123 680	1965	1 638 000
1976	33 727	1979	140 136		

联邦政府正在考虑对超过 5 000 000 的飓风损失量进行 100%赔付的资金计划. 请你对此计划进行初步的估计:

(a) 估计飓风造成损失的平均值、标准差、变异系数 (coefficient of variation) 以及偏度;

(b) 估计限额为 500 000 000 时损失的一阶有限矩和二阶有限矩.

10.3* 现有 30 个索赔额的随机采样结果, 其中有 2 个索赔额为 2 000, 6 个为 4 000, 12 个为 6 000, 10 个为 8 000. 求经验分布的偏度系数.

10.3 分组数据的经验分布

对于分组数据, 根据之前的定义构造经验分布函数是不可能的. 不过, 仍然有可能近似估算经验分布函数. 其方法是在任何可能的点计算经验分布函数的值, 再用某种合理的方式将这些值联结起来. 对于分组数据, 常将数据点用直线连接近似计算分布函数, 其余的插值方法将在第 15 章中讨论. 以下将各个组的分界点记为 $c_0 < c_1 < \cdots < c_k$, 其中 $c_0 = 0$ 且 $c_k = \infty$. 落在 c_{j-1} 和 c_j 之间的观测值的个数记为 n_j, 则有 $\sum\limits_{j=1}^{k} n_j = n$. 对这样的数据, 经验分布函数在各组分界点处的值可以直接计算, 即 $F_n(c_j) = \frac{1}{n}\sum\limits_{i=1}^{j} n_i$. 注意到这里并没有对落在区间分间界点上的观测值做出任何规定. 事实上, 对此并没有一个公认正确的处理方法, 但是不管采用何种方法, 在处理每个区间时应当使用一致的方法. 观察数据集 C 可以发现, 根本无法

得知它采用什么方法来记录分界点上的数据. 如果我们对这方面的知识有所了解, 就不会影响到接下来的计算.[①]

定义 10.12 对于分组数据, 将各个组分界点处的经验分布函数值用直线连接起来得到的经验分布函数的估计值叫做卵形线 (ogive) 其计算公式如下

$$F_n(x)=\frac{c_j-x}{c_j-c_{j-1}}F_n(c_{j-1})+\frac{x-c_{j-1}}{c_j-c_{j-1}}F_n(c_j),\quad c_{j-1}\leqslant x\leqslant c_j.$$

除了组间的分界点, 该函数在其余任何点都可导. 因此, 为了完整地描述密度函数, 只需人为地将其定义为右连续的函数即可.

定义 10.13 对于分组数据, 对卵形线求导得到经验密度函数. 由此得到的密度函数被称为直方图 (histogram). 其计算公式为

$$f_n(x)=\frac{F_n(c_j)-F_n(c_{j-1})}{c_j-c_{j-1}}=\frac{n_j}{n(c_j-c_{j-1})},\quad c_{j-1}\leqslant x<c_j.$$

很多计算机程序生成的直方图其实只是柱状图, 柱的高度与 n_j/n 成比例. 当分组的区间长度相同时, 这种方法是可以接受的; 但如果区间长度不同, 就必须用上面的公式计算. 这种方法的最大优点在于, 直方图确实代表密度函数, 并且直方图下方图形的面积可以用来得到经验概率值.

例 10.14 构造数据集 C 的卵形线和直方图.

解 分布函数为

$$F_{227}(x)=\begin{cases}0.000\,058\,150x, & 0\leqslant x\leqslant 7\,500,\\ 0.297\,36+0.000\,018\,502x, & 7\,500\leqslant x\leqslant 17\,500,\\ 0.472\,10+0.000\,008\,517x, & 17\,500\leqslant x\leqslant 32\,500,\\ 0.634\,36+0.000\,003\,524x, & 32\,500\leqslant x\leqslant 67\,500,\\ 0.784\,33+0.000\,001\,302x, & 67\,500\leqslant x\leqslant 125\,000,\\ 0.918\,82+0.000\,000\,227x, & 125\,000\leqslant x\leqslant 300\,000,\\ \text{未定义}, & x>300\,000,\end{cases}$$

其中的计算如下所示, 例如对 $32\,500\leqslant x\leqslant 67\,500$

$$F_{227}(x)=\frac{67\,500-x}{67\,500-32\,500}\frac{170}{227}+\frac{x-32\,500}{67\,500-32\,500}\frac{198}{227}.$$

① 从技术角度讲, 为了使 $F_n(c_j)$ 成为经验分布函数, 对于从 c_{j-1} 到 c_j 的区间. 应当包含 $x=c_j$ 而不应包含 $x=c_{j-1}$.

对超过 300 000 的部分, 分布函数无定义, 因为最后一个区间的宽度为无穷大. 图 10-1 描绘了在 0 到 125 000 之间卵形线的图形

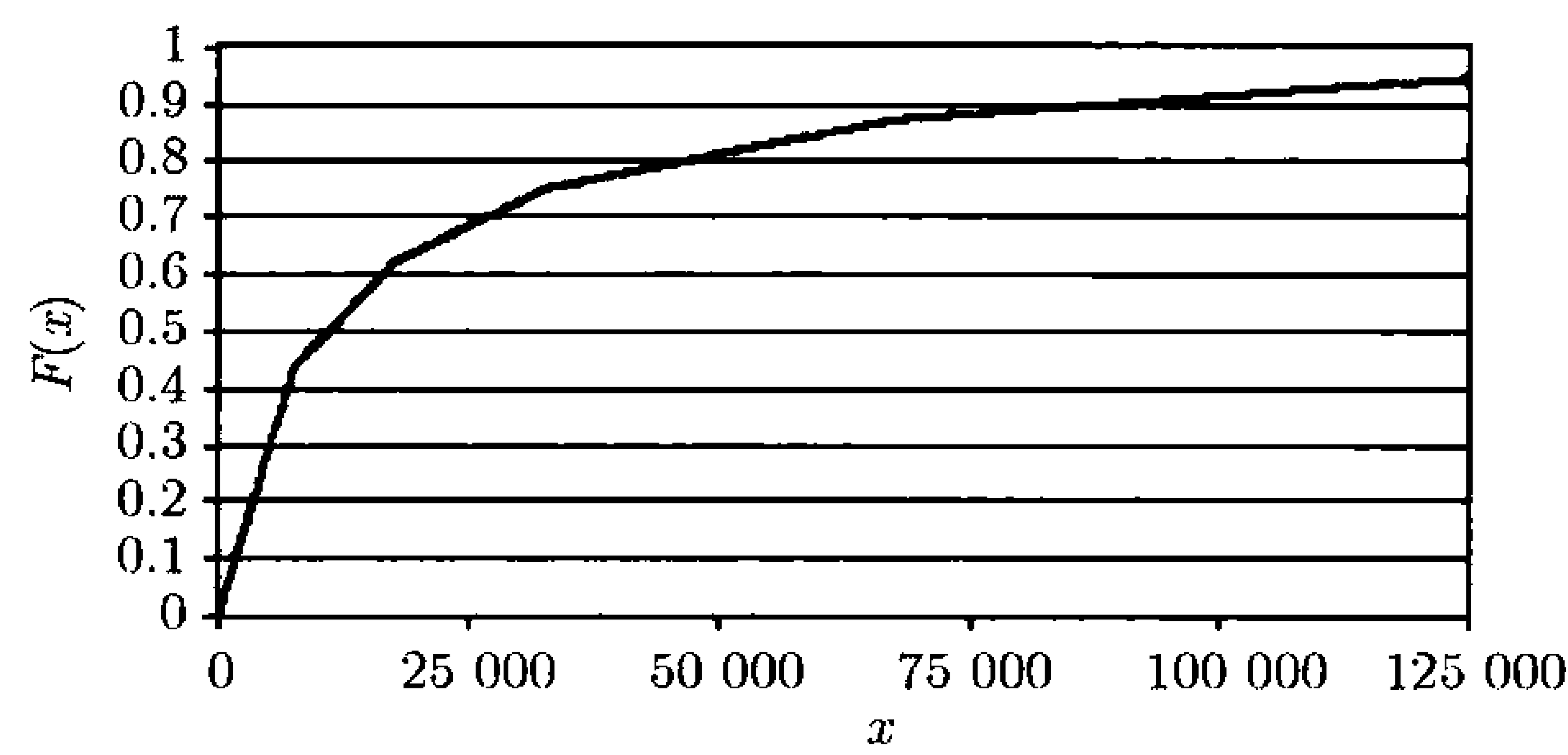


图 10-1　非寿险赔付额的卵形线

其导数为如下所示的阶梯函数

$$f_{227}(x)=\begin{cases}0.000\ 058\ 150, & 0\leqslant x<7\ 500,\\ 0.000\ 018\ 502, & 7\ 500\leqslant x<17\ 500,\\ 0.000\ 008\ 517, & 17\ 500\leqslant x<32\ 500,\\ 0.000\ 003\ 524, & 32\ 500\leqslant x<67\ 500\\ 0.000\ 001\ 302, & 67\ 500\leqslant x<125\ 000,\\ 0.000\ 000\ 227, & 125\ 000\leqslant x<300\ 000,\\ \text{未定义}, & x\geqslant 300\ 000.\end{cases}$$

这个函数在 0 到 125 000 的图形如图 10-2 所示.　□

表 10-9　习题 10.4 的数据

赔付额范围	赔付笔数
0～25	6
25～50	24
50～75	30
75～100	31
100～150	57
150～250	80
250～500	85
500～1000	54
1000～2000	15
2000～4000	10
4 000 以上	0

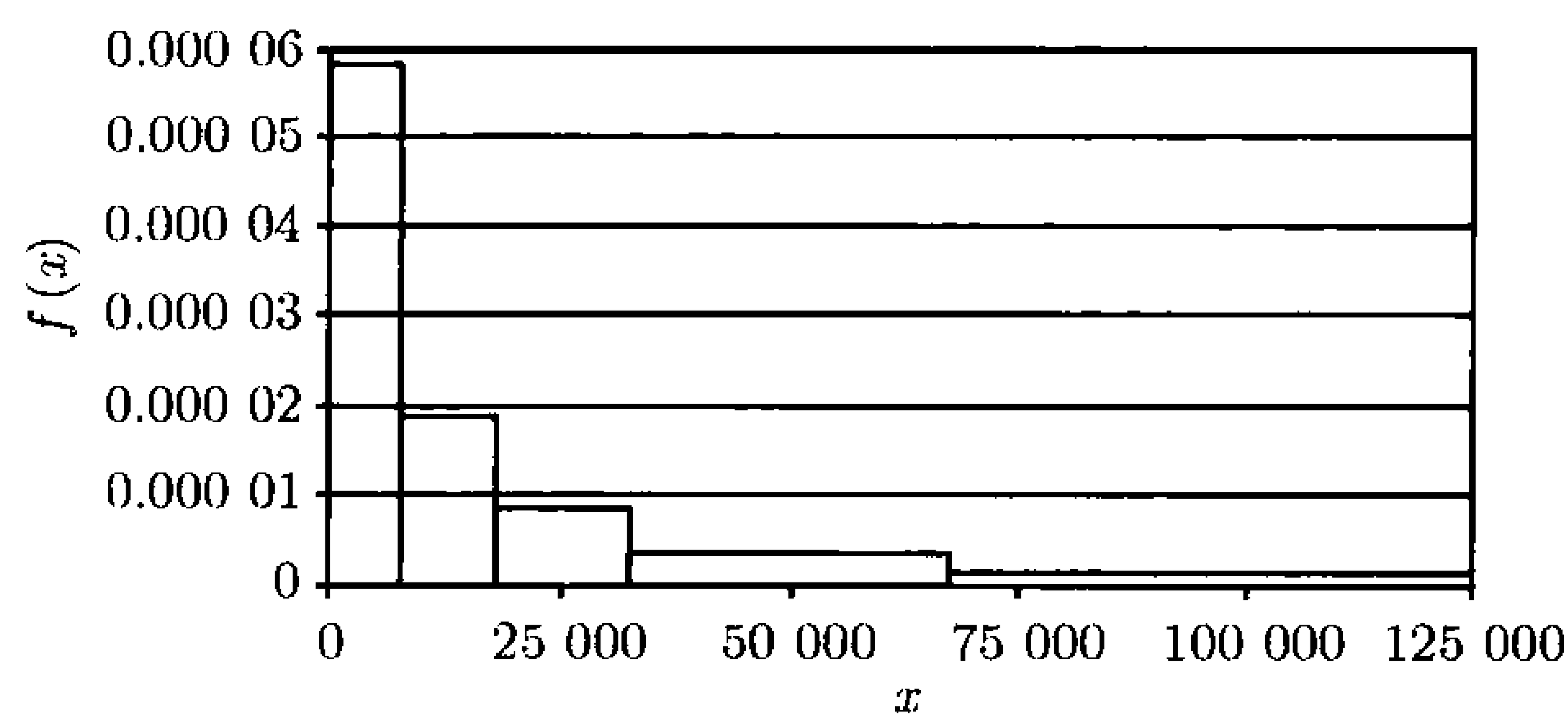


图 10-2 非寿险赔付额的矩形线

习题

10.4 构造表 10-9 中数据的卵形线和直方图.

10.5* 以下 20 个数据是某年记录的 20 次风灾的损失额 (单位为百万美元):

1	1	1	1	1	2	2	3	3	4
6	6	8	10	13	14	15	18	22	25

(a) 构造一个卵形线, 用 0.5, 2.5, 8.5, 15.5 和 29.5 作为分界点.

(b) 运用 (a) 中的分界点构造直方图.

10.6 表 10-10 中的数据来自 Herzog and Laverty[54], 记录了一组截至 1993 年 12 月 31 日的 15 年的抵押贷款的情况. 这组贷款按照其发放时的状况分为 2 类: 一类是已有贷款的再融资, 一类是初次发放的贷款. 表中的数据显示了贷款发放的总数和在指定年数之后仍然有效的贷款占总数的百分比. 请根据已知的数据尽可能精确地画出 2 条卵形线 (在同一个图内). 观察卵形线的形状, 判断初次发放的贷款和再融资贷款的生存时间 (即贷款从发放到终结的时间) 的分布是否存在明显的不同?

表 10-10 习题 10.6 的数据

年数	再筹资		初次发放	
	发放数量	生存数量	发放数量	生存数量
1.5	42 300	99.97	12 813	99.88
2.5	9 756	99.82	18 787	99.43
3.5	1 550	99.03	22 513	98.81
4.5	1 256	98.41	21 420	98.26
5.5	1 619	97.78	26 790	97.45

10.7* 根据表 10-11 中收集的数据构造直方图.

表 10-11 习题 10.7 的数据

损失额	观察值个数
0~2	25
2~10	10
10~100	10
100~1 000	5

10.8* 现有 40 个损失额的观测值. 其中有 16 个数值位于 1 到 $\frac{4}{3}$ 之间, 其总和为 20; 有 10 个数值位于 $\frac{4}{3}$ 到 2 之间, 总和为 15; 有 10 个数值在 2 到 4 之间, 总和为 35; 剩下 4 个损失额大于 4. 基于这些观测数据, 用经验模型计算 $\mathrm{E}(X \wedge 2)$.

10.9* 一个容量为 2 000 的样本包含 1 700 个不大于 6 000 的观测值, 30 个大于 6 000 但不大于 7 000 的观测值, 以及 270 个大于 7 000 的观测值. 位于 6 000 到 7 000 之间的 30 个观测值的总和是 200 000. 基于上述数据, 用经验分布计算 $\mathrm{E}(X \wedge 6\,000)$ 为 1 810. 请用经验分布计算 $\mathrm{E}(X \wedge 7\,000)$.

第 11 章　基于修正数据的统计估计

11.1　点　估　计

由于存在删失和截断, 数据的不完整现象很常见, 其严格定义如下.

定义 11.1　*若对观测值小于 d 的数据不作记录, 而观测值大于 d 时如实记录, 则称这个数据在 d下截断(也称左截断).*

若对观测值大于 u 的数据不作记录, 而观测值小于 u 时如实记录, 则称这个数据在 u上截断(也称右截断).

若对观测值小于 d 的数据记录为 d, 而观测值大于 d 时如实记录, 则称这个数据在 d下删失(也称左删失).

若对观测值大于 u 时的数据记录为 u, 而观测值小于 u 时如实记录, 则称这个数据在 u上删失(也叫右删失).

最常见的情形是左截断和右删失. 很多保单都有一个免赔额 d, 这时就自然出现了左截断 —— 如果投保人的损失额低于 d, 他会意识到自己不会得到任何赔付, 因此也就不会告知保险人自己的损失情况; 相反, 如果损失额高于 d, 则保险人就会得到关于这个损失额的报告. 限额保险是右删失的例子. 如果损失额超过了 u, 超过部分不会由保险人赔付, 超过部分也就没有相应的记录; 但是, 发生了一个金额至少为 u 的损失这个事实被记录下来.

在生命表构造中, 要跟踪观测每个人的出生和身故时间是不现实的. 更为常见的做法是: 在几年的时间内, 跟踪观测一个由不同年龄段的人组成的人群的生存状况. 当某人开始参与一项生存研究项目时, 他必然处于生存状态, 且其身故年龄至少不会低于他参加该项目时的年龄, 因此数据是左截断的. 如果这个参与者在研究项目结束时还未身故, 就出现了右删失的现象, 具体的身故年龄无法确定, 但可以确定的是这个身故年龄不会小于此人在项目结束时的年龄. 当有人由于退保等原因在项目结束前退出时, 右删失也会发生.

因为左截断和右删失是精算工作中最常遇到的情形, 所以本章只讨论这 2 种情形. 为简明起见, 左截断也简称*截断*, 右删失也简称*删失*.

为了用删失或者截断的数据构造经验分布函数, 第一个任务是明确一些记号来表示与数据信息相关的量. 对于个体数据, 有 3 个因素是必需的. 首先是数据观测值的截断点 — 用 d_j 表示第 j 个观测值的截断点, 如果没有截断发生, 则令 $d_j = 0$. 其次是观测值本身 — 根据数据是否删失, 其表示符号有所不同: 如果该数据不是

删失的数据, 将其值记为 x_j, 否则记为 u_j. 如果更加严格地描述这个问题, 删失点在事先已知和未知的情形下是有区别的. 例如, 限额保险的删失通常都是在索赔发生之前就确定了; 但在另一种以被保险人的寿命记录为基础进行死亡率的研究中, 当保单出售时, 未来的退保年龄作为一个删失点是未知的. 本章不考虑这 2 种情形的差别.

为了进行估计, 原始数据必须按照惯例模式进行标准化处理. 最值得注意的数据是那些没有删失的观测值. 用 $y_1 < \cdots < y_k$ 表示样本中出现的 k 个互不相同的 x_j 的值, 用 s_j 表示这些没有删失的观测值 y_j 在样本中出现的次数. 最后一个重要的量是第 j 个观测值 y_j 处的**风险集**, 记为 r_j. 在研究和死亡力有关的问题时, 风险集由那些在指定的年龄仍处在被观察状态的个体构成, 包括所有在某年龄 (用 x 表示) 或者之后身故的个体, 以及所有身故时间观测值在某年龄 (用 u 表示) 或者之后删失的个体. 但是, 对那些在某年龄 (用 d 表示) 或之后首次观测到的个体, 我们认为其在指定的时点并没有处于被观测的状态. 计算公式如下:

$$r_j = (\text{大于等于}y_j\text{的}x_i\text{的个数})+(\text{大于等于}y_j\text{的}u_i\text{的个数})-(\text{大于等于}y_j\text{的}d_i\text{的个数}).$$

另外, 考虑到 d_i 的总个数同 x_i 及 u_i 的总个数之和相等, 因此 r_j 也可以由以下公式计算

$$r_j = (\text{小于}y_j\text{的}d_i\text{的个数}) - (\text{小于}y_j\text{的}x_i\text{的个数}) - (\text{小于}y_j\text{的}u_i\text{的个数}). \tag{11.1}$$

在这 2 个公式中, 后者更容易从直观上理解, 因为它是由所有在给定年龄之前参与研究的个体数减去已经离去的个体数. 最关键的一点在于, 风险集是所有在年龄 y_j 仍然生存的个体总数. 如果数据是损失额, 风险集就是损失观测个数 (实际金额或者由保单限额决定的最大金额) 大于或等于 y_j 的保单数减去免赔额大于或等于 y_j 的保单数. 还可以得到这个公式的递归形式

$$\begin{aligned} r_j = r_{j-1} &+ (\text{在}y_{j-1}\text{和}y_j\text{之间的}d_i\text{个数}) - (\text{等于}y_{j-1}\text{的}x_i\text{个数}) \\ &-(\text{在}y_{j-1}\text{和}y_j\text{之间的}u_i\text{个数}), \end{aligned} \tag{11.2}$$

其中 "**之间**" 是指大于或等于 y_{j-1} 且小于 y_j, 并规定 $r_0 = 0$.

例 11.2 对数据集 D2 同时利用公式 (11.1) 和公式 (11.2) 计算以上定义的这些量.

解 计算结果列于表 11-1 和表 11-2. □

尽管到目前为止在这个问题上我们已经做了不少讨论, 但给出生存函数的估计方法才是我们必须解决的问题. 最常用的估计方法为Kaplan-Meier **有限乘积估计法**(Kaplan-Meier product-limit Estimator)[70]. 由于在 y_1 年龄之前没有人身故, 生存函数在这个年龄之前恒等于 1, 所以将初始值取为 $S(0) = 1$. 考虑到在此条件

下, 在 y_1 年龄之前, 共有 r_1 人可能会身故, 将确实身故的人数记为 s_1. 这样一来, 活过 y_1 年龄的概率就是 $(r_1-s_1)/r_1$, 这就是 $S(y_1)$ 的值, 并且生存函数在 y_2 之前将保持这个值不变. 继续以上的推导可得知这时的生存函数在 y_2 点的新的取值是 $S(y_1)(r_2-s_2)/r_2$. 一般公式为

$$S_n(t)=\begin{cases}1, & 0\leqslant t<y_1,\\ \prod_{i=1}^{j=1}\left(\dfrac{r_i-s_i}{r_i}\right), & y_{j-1}\leqslant t<y_j, j=2,\cdots,k,\\ \prod_{i=1}^{k}\left(\dfrac{r_i-s_i}{r_i}\right)\text{或}\,0, & t\leqslant y_k.\end{cases}$$

表 11-1 例 11.2 的数据

i	d_i	x_i	u_i	i	d_i	x_i	u_i
1	0	–	0.1	16	0	4.8	–
2	0	–	0.5	17	0	–	4.8
3	0	–	0.8	18	0	–	4.8
4	0	0.8	–	19~30	0	–	5.0
5	0	–	1.8	31	0.3	–	5.0
6	0	–	1.8	32	0.7	–	5.0
7	0	–	2.1	33	1.0	4.1	–
8	0	–	2.5	34	1.8	3.1	–
9	0	–	2.8	35	2.1	–	3.9
10	0	2.9	–	36	2.9	–	5.0
11	0	2.9	–	37	2.9	–	4.8
12	0	–	3.9	38	3.2	4.0	–
13	0	4.0	–	39	3.4	–	5.0
14	0	–	4.0	40	3.9	–	5.0
15	0	–	4.1				

表 11-2 例 11.2 的风险集计算

j	y_j	s_j	r_j
1	0.8	1	32−0−2=30 或 0+32−0−2=30
2	2.9	2	35−1−8=26 或 30+3−1−6=26
3	3.1	1	37−3−8=26 或 26+2−2−0=26
4	4.0	2	40−4−10=26 或 26+3−1−2=26
5	4.1	1	40−6−11=23 或 26+0−2−1=23
6	4.8	1	40−7−12=21 或 23+0−1−1=21

如果 $s_k=r_k$, 则当 $t\geqslant y_k$ 时 $S(t)=0$ 才有意义. 该样本中的每个个体都在 y_k 年龄之前身故, 因此从这组经验看, 活过这个年龄是不可能的. 但是, 由于存在删失, 可能在最后一个个体身故的时候, 仍有个体生存, 不过这些生存个体都在此年

龄之前删失了, 其实活过最后一个身故者的身故年龄是有可能的, 但是没有经验数据来构造这个年龄之后的生存函数. 一种解决办法 (上述公式中最后一行的第一个选择) 是取最后得到的函数值, 这无疑是最合理的选择. 另一种解决办法是将函数在最后一个身故年龄之后的取值都规定为零, 无论这个身故年龄是一个实际的身故年龄还是一个删失的年龄, 这种方法有着基本的合理性并且可以用之计算各阶矩. 一种折中的选择是用一条指数衰减的曲线让函数值由当前的取值逐渐趋于零. 令 $w=\max\{x_1,\cdots,x_n,u_1,\cdots,u_n\}$, 则对 $t\geqslant w$ 有

$$S_n(t)=\mathrm{e}^{(t/w)\ln s^*}=(s^*)^{t/w},\text{其中 } s^*=\prod_{i=1}^{k}\left(\frac{r_i-s_i}{r_i}\right).$$

如果使用 Excel ®工作表[①]计算 s_j 和 r_j 的值, 将更为快捷, 具体步骤如下.

(1) 每个数据观测点为工作表的一行输入, 数据点之间无须按照特别的顺序排列.

(2) 每行有 3 项输入：第 1 项为 d_j, 第 2 项为 x_j 或 u_j, 第 3 项应根据第 2 项的内容输入一个字母 —— 当第 2 项是 x 值 (即实际观测值) 时, 在第 3 项中输入字母 "x"(不带这里的引号); 当第 2 项是 u 值 (即删失值) 时, 在第 3 项中输入字母 "u". 例如, 假设单元格 B6:B45 为 d_j, 则 C6:C45 为 x_j 或 u_j, D6:D45 则为字母 x 或 u.

(3) 按照下述方法创建计算 x, r 和 s 的列.

(4) 为了将观测值排序, 在单元格 F2 输入最小值 d, 即公式 MIN(B6:B45).

(5) 在单元格 F3 中输入公式

$$=\mathrm{MIN(IF(C\$6:C\$45>F2,IF(D\$6:D\$45=\text{“x”},C\$6:C\$45,1E36),1E36))}.$$

由于这是一个数组公式, 它必须由 Ctrl+Shift+Enter 完成输入. 将这个公式复制到单元格 F4, 单元格 F5, ⋯⋯ 直到出现数值 E36. 现在单元格 F 列显示的是没有重复的有序的 y 值.

(6) 在单元格 G3 中输入公式

$$=\mathrm{COUNTIF(B\$6:B\$45,\text{“ < ”}\&F3)-COUNTIF(C\$6:C\$45,\text{“ < ”}\&F3)}.$$

将此公式复制到单元格 G4, 单元格 G5, ⋯⋯ 直到单元格的左边是 F 列的倒数第二个非空的单元格. 这一列的各个数值就是风险集的值.

---

① 这一系列计算步骤是由 Charles Thayer 设计并由 Margie Rosenberg 改进的, 是对作者的早期版本计算方法的重大改进. 这里所列的语句为 Office XP 形式, 其他版本 Office 软件的计算步骤可与之相似.

(7) 在单元格 H3 中输入公式

$$= \text{SUM}(\text{IF}(\text{C\$6}:\text{C\$45}=\text{F3},\text{IF}(\text{D\$6}:\text{D\$45}=\text{“x”},1,0),0)).$$

用 Ctrl+Shift+Enter 输入这个公式, 并将其复制到单元格 H4, 单元格 H5, …… 直到与 G 列的行数相同. 这一列显示的是 s 的值.

(8) 在单元格 I2 中输入 1, 开始计算 $S(y)$ 的值.

(9) 在单元格 I3 中输入以下公式就可以得到 $S(y)$ 的下一个取值

$$= 12^*(\text{G3}-\text{H3})/\text{G3}.$$

然后将这个公式复制到单元格 I4, 单元格 I5, …… 直到完成计算.

例 11.3 求数据集 D2 的 Kaplan-Meier 估计.

解 由上例的计算, 有

$$S_{40}(t)=\begin{cases}1, & 0\leqslant t<0.8,\\ \frac{30-1}{30}=0.966\,7, & 0.8\leqslant t<2.9,\\ 0.966\,7\frac{26-2}{26}=0.892\,3, & 2.9\leqslant t<3.1,\\ 0.892\,3\frac{26-1}{26}=0858\,0, & 3.1\leqslant t<4.0,\\ 0.858\,0\frac{26-2}{26}=0.792\,0, & 4.0\leqslant t<4.1,\\ 0.792\,0\frac{23-1}{23}=0.757\,6, & 4.1\leqslant t<4.8,\\ 0.757\,6\frac{21-1}{21}=0.721\,5, & 4.8\leqslant t<5.0,\\ 0.721\,5 \text{ 或 } 0 \text{ 或 } 0.721\,5^{t/5.0}, & t\geqslant 5.0.\end{cases}$$ □

除了 Kaplan-Meier 估计外, 还可以用前文介绍的 Nelson- Åalen 方法的修正进行估计. 正如前文所述, 这种修正的 Nelson- Åalen 估计法可以直接估计累积风险率函数. 下面将给出这种估计的直观推导. 对任意时刻 t, 令 $r(t)$ 是风险集, $h(t)$ 是风险率函数, $s(t)$ 仍表示在时刻 t 之前身故的个体数目的期望值, 则可得以下结论

$$s(t)=\int_0^t r(u)h(u)\mathrm{d}u.$$

两边求导, 有

$$\mathrm{d}s(t)=r(t)h(t)\mathrm{d}t.$$

于是

$$\frac{\mathrm{d}s(t)}{r(t)}=h(t)\mathrm{d}t.$$

两边积分, 得

$$\int_0^t \frac{\mathrm{d}s(u)}{r(u)} = \int_0^t h(u)\mathrm{d}u = H(t).$$

现在将真实的期望值 $s(t)$ 用 t 时刻之前的身故个体数的观测值 $\hat{s}(t)$ 代替. $\hat{s}(t)$ 是一个阶梯函数, 在每一个身故发生的时刻增加 s_i, 因此上式的左边变为

$$\sum_{t_i \leqslant t} \frac{s_i}{r_i},$$

这就给出了估计值 $\hat{H}(t)$. 则 Nelson- Åalen 估计为

$$\hat{H}(t) = \begin{cases} 0, & 0 \leqslant t < y_1, \\ \displaystyle\sum_{i=1}^{j-1} \frac{s_i}{r_i}, & y_{j-1} \leqslant t < y_j, \quad j = 2, \cdots, k, \\ \displaystyle\sum_{i=1}^{k} \frac{s_i}{r_i}, & t \geqslant y_k, \end{cases}$$

进而, 有

$$\hat{S}(t) = \mathrm{e}^{-\hat{H}(t)}.$$

对 $t \geqslant w$, 也可以用 $\hat{S}(t) = 0$ 或 $\hat{S}(t) = \hat{S}(y_k)^{t/w}$ 进行估计.

例 11.4　求数据集 D2 生存函数的 Nelson-Åalen 估计.

解

$$\hat{H}(t) = \begin{cases} 0, & 0 \leqslant t < 0.8, \\ \frac{1}{30} = 0.0333, & 0.8 \leqslant t < 2.9, \\ 0.0333 + \frac{2}{26} = 0.1103, & 2.9 \leqslant t < 3.1, \\ 0.1103 + \frac{1}{26} = 0.1487, & 3.1 \leqslant t < 4.0, \\ 0.1487 + \frac{2}{26} = 0.2256, & 4.0 \leqslant t < 4.1, \\ 0.2256 + \frac{1}{23} = 0.2691, & 4.1 \leqslant t < 4.8, \\ 0.2691 + \frac{1}{21} = 0.3167, & t \geqslant 4.8. \end{cases}$$

$$\hat{S}(t) = \begin{cases} 1, & 0 \leqslant t < 0.8, \\ \mathrm{e}^{-0.0333} = 0.9672, & 0.8 \leqslant t < 2.9, \\ \mathrm{e}^{-0.1103} = 0.8956, & 2.9 \leqslant t < 3.1, \\ \mathrm{e}^{-0.1487} = 0.8618, & 3.1 \leqslant t < 4.0, \\ \mathrm{e}^{-0.2256} = 0.7980, & 4.0 \leqslant t < 4.1, \\ \mathrm{e}^{-0.2691} = 0.7641, & 4.1 \leqslant t < 4.8, \\ \mathrm{e}^{-0.3167} = 0.7285, & 4.8 \leqslant t < 5.0, \\ 0.7285 \text{ 或 } 0 \text{ 或 } 0.7285^{t/5.0}, & t \geqslant 5.0. \end{cases}$$

□

需要特别注意的是, 如果数据是截断的, 得到的分布函数其实是在给定赔付额超过最小截断点 (即最小的 d 值) 的条件下的赔付额分布. 从经验上讲, 由于没有任何关于小于这个最小截断点的观测值的信息, 因此也就没有分布函数在此范围内的任何信息. 另外应当注意的是, 这一节采用的所有记号和公式同 10.2 节是完全一致的, 如果实际应用中没有出现删失和截断数据, 那么, 用本节的公式得到的结果和用 10.2 节中的经验公式得到的结果应该是相同的.

习题

11.1 将 "退保" 看作 "身故", 重新计算例 11.2. 最简明的办法是交换字母 x 和 u, 然后利用前文提到的公式计算. 在这种情形下, 身故其实是产生删失的因素, 因为我们无法对身故者继续观察, 导致其退保时间无从知晓. 将生存至 5 年保单期满且未退保的个体看作是在 5 年期末退保的.

11.2 求数据集 D2 的退保时间的 Kaplan-Meier 估计. 将生存至 5 年保单期满且未退保的个体看作是在 5 年期末退保的.

11.3 使用数据集 D2 中的数据, 以退保时间为变量, 计算 $H(t)$ 和 $S(t)$ 的 Nelson-Åalen 估计.

11.4 运用 Kaplan-Meier 估计法和 Nelson-Åalen 估计法, 计算工伤险损失额的分布函数. 首先用数据集 B 中的原始数据计算, 然后将数据修正为左截断点为 100 和右删失点为 1 000, 重新计算.

11.5* 已知从某车险保单中随机选出的 5 份保单的首次索赔时间分别为: 1, 2, 3, 4, 5. 其后被告之其中的某一个是退保, 但不知具体为哪一个. 试问上述哪个首次索赔为退保保单时, 由有限乘积方法得到 $S(4)$ 的最小估计就是该首次索赔时间在 4 之后的概率?

11.6* 在一项关于死亡力的研究中, 数据为右删失的, 已知信息由表 11-3 所示. 用 Nelson-Åalen 方法计算生存函数在 12 的估计值.

表 11-3　习题 11.6 的数据

时间 t_j	身故个体数目 s_j	风险数目 r_j
5	2	15
7	1	12
10	1	10
12	2	6

11.7* 现有 300 只老鼠, 从出生时开始观察, 另有 20 只老鼠从年龄为 2(天) 时开始观察, 还有 30 只老鼠从年龄为 4 时开始观察. 在所有这些老鼠中, 有 6 只在年龄为 1 时死亡, 10 只在年龄为 3 时死亡, 10 只在年龄为 4 时死亡, a 只在年龄为 5 时死亡, b 只在年龄为 9 时死亡, 6 只在年龄为 12 时死亡. 另外, 有 45 只老鼠在年龄为 7 时 (因逃跑、丢失等原因) 失去观察, 35 只在年龄为 10 时失去观察, 15 只在年龄为 13 时失去观察.

已知由有限乘积方法估计如下结果: $S_{350}(7) = 0.892$, $S_{350}(13) = 0.856$, 求 a 和 b 的值.

11.8* 令 n 为从出生开始观测的个体数目, 观测过程中没有删失现象且任何两个个体都没有在同一年龄身故. 在第 9 次身故事件发生时, 用 Nelson-Åalen 方法估计的累积风险率为 0.511, 且在第 10 次身故事件发生时, 这个数值变为 0.588. 试估计在第 3 次身故事件发生时的生存函数值.

11.9* 现有一研究项目, 其研究对象都是从出生时就开始观测的个体. 但是, 一些研究对象可能因为身故之外的原因离开该研究项目. 在第 3 个年龄段有 1 人身故 (即 $s_3 = 1$); 在第 4 个年龄段有 2 人身故; 在第 5 个年龄段有 1 人身故. 已知用有限乘积方法估计如下结果: $S_n(y_3) = 0.72$, $S_n(y_4) = 0.60$ 和 $S_n(y_5) = 0.50$. 假设没有观测值在身故事件发生的时刻删失, 求在 y_4 时刻和 y_5 时刻之间删失的观测值数目.

11.2　均值、方差以及置信区间的估计

如果所有信息都可以得到, 则可以直接对生存函数的经验估计进行相关的计算.

例 11.5　证明: 基于完整数据对分布函数进行的经验估计是无偏且相合的.

证明　$S(x)$ 的经验估计为 $S_n(x) = Y/n$, 其中 Y 为样本中大于 x 的观测数目, 因此 Y 服从参数为 n 和 $S(x)$ 的二项分布, 此时, 有

$$\mathrm{E}[S_n(x)] = \mathrm{E}\left(\frac{Y}{n}\right) = \frac{nS(x)}{n} = S(x),$$

证明了此估计是无偏的. 方差可由下式计算

$$\mathrm{Var}[S_n(x)] = \mathrm{Var}\left(\frac{Y}{n}\right) = \frac{S(x)[1 - S(x)]}{n},$$

此式极限为零, 这就证明了相合性. □

可以利用以上结果得到方差的估计. 往往并不知道 $S(x)$ 的真实值, 因为这正是我们要估计的量. 方差的估计值由下式给出

$$\widehat{\mathrm{Var}}[S_n(x)] = \frac{S_n(x)[1 - S_n(x)]}{n}.$$

当使用经验估计时, 这个结果仍然成立. 令 $p = \Pr(a < X \leqslant b)$, 则 p 的经验估计值为 $\hat{p} = S_n(a) - S_n(b)$. 与上例的推导类似, 可以证明 $\hat{p}$ 是无偏且相合的, 且有 $\mathrm{Var}(\hat{p}) = p(1 - p)/n$.

在研究死亡力或者在估计免赔额的影响时, 常常更关心一些条件量的估计.

例 11.6　对数据集 D1 的所有观测信息, 用经验方法估计 q_2 及该估计值的方差.

解 对于这个数据集, $n=30$, 在时刻 2 之前有 1 人身故, 在时刻 3 之前有 3 人身故, 因此有 $S_{30}(2)=\frac{29}{30}$ 以及 $S_{30}(3)=\frac{27}{30}$, 经验估计为

$$\widehat{q2}=\frac{S_{30}(2)-S_{30}(3)}{S_{30}(2)}=\frac{2}{29}.$$

本题的难点在于计算这个估计量的平均值和方差. 用 X 表示在 0 到 2 之间身故的人数, Y 表示在 2 到 3 之间身故的人数, 则 $\hat{q}_2=Y/(30-X)$. 很明显 $\mathrm{E}(\hat{q}_2)$ 实际上是无法计算的, 因为 X 恰好取值为 30 的概率为一个整数, 此时这个估计量是没有意义的①. 通常的解决办法是计算这个估计量的条件方差, 即在给定有 29 人在 2 时刻存活的条件下求方差. 此时唯一的随机变量是 Y, 进而有

$$\widehat{\mathrm{Var}}\left[\widehat{q2}|S_{30}(2)=\frac{29}{30}\right]=\frac{(2/29)(27/29)}{29}.$$ □

一般令 n 为最初的样本容量, n_x 为年龄 x 仍然存活的个体数, n_y 为年龄 y 仍然存活的个体数, 则有

$$\widehat{\mathrm{Var}}({}_{y-x}\hat{q}_x|n_x)=\widehat{\mathrm{Var}}({}_{y-x}\hat{p}_x|n_x)=\frac{(n_x-n_y)(n_y)}{n_x^3}.$$

例 11.7 对数据集 B 假设免赔额为 250, 用经验方法估计赔付额不低于 1 000 的概率.

解 通过经验分析发现有 13 个损失额在免赔额之上, 其中有 4 个超过了 1 250(对这些损失的赔付为 1 000), 故经验估计是 4/13, 用生存函数的记号可表示为 $S_{20}(1\,250)/S_{20}(250)$. 这里又一次强调, 关于这个估计量只有条件方差可以通过计算求得, 这个方差的估计值是 $4(9)/13^3$. □

对于分组数据, 生存函数在组间分界点处的估计不是问题, 对这些分界点之间的部分应使用卵形线进行插值计算, 其过程稍显复杂.

例 11.8 用卵形线计算生存函数的估计量的期望值和方差; 用直方图计算密度函数估计量的期望值和方差.

解 假设 x 介于分界点 c_{j-1} 和 c_j 之间, 用 Y 表示在 c_{j-1} 以下的观测值的数目, 用 Z 表示在 c_{j-1} 以上但不超过 c_j 的观测值的数目, 则

$$S_n(x)=1-\frac{Y(c_j-c_{j-1})+Z(x-c_{j-1})}{n(c_j-c_{j-1})},$$

① 在这种情形下贝叶斯方法 (将在 12.4 节中介绍) 更为有效. 贝叶斯决策仅仅依赖数据的观测值进行推断, 而和其他可能的观测量没有关系. 如果 $X=30$, 没有继续估计的必要, 而如果 $X<30$, 分析可以继续进行.

且

$$\begin{aligned}\mathrm{E}[S_n(x)] &= 1 - \frac{n[1-S(c_{j-1})](c_j - c_{j-1}) + n[S(c_{j-1}) - S(c_j)](x - c_{j-1})}{n(c_j - c_{j-1})} \\ &= S(c_{j-1})\frac{c_j - x}{c_j - c_{j-1}} + S(c_j)\frac{x - c_{j-1}}{c_j - c_{j-1}}.\end{aligned}$$

这个估计是有偏的 (尽管它是真实插值的无偏估计), 其方差为

$$\mathrm{Var}[S_n(x)] = \frac{\begin{array}{c}(c_j - c_{j-1})^2\mathrm{Var}(Y) + (x - c_{j-1})^2\mathrm{Var}(Z) \\ +2(c_j - c_{j-1})(x - c_{j-1})\mathrm{Cov}(Y, Z)\end{array}}{[n(c_j - c_{j-1})]^2},$$

其中 $\mathrm{Var}(Y) = nS(c_{j-1})[1-S(c_{j-1})], \mathrm{Var}(Z) = n[S(c_{j-1})-S(c_j)][1-S(c_{j-1})+S(c_j)]$, 且 $\mathrm{Cov}(Y, Z) = -n[1 - S(c_{j-1})][S(c_{j-1}) - S(c_j)]$. 密度函数的估计量为

$$f_n(x) = \frac{Z}{n(c_j - c_{j-1})},$$

且

$$\mathrm{E}[f_n(x)] = \frac{S(c_{j-1}) - S(c_j)}{c_j - c_{j-1}},$$

这是真实的密度函数的有偏估计, 其方差为

$$\mathrm{Var}[f_n(x)] = \frac{[S(c_{j-1}) - S(c_j)][1 - S(c_{j-1}) + S(c_j)]}{n(c_j - c_{j-1})^2}. \qquad \square$$

例 11.9 基于数据集 C 估计 $S(10\ 000)$ 和 $f(10\ 000)$ 以及这两个估计值的方差.

解 点估计为

$$\begin{aligned}S_{227}(10\ 000) &= 1 - \frac{99(17\ 500 - 7\ 500) + 42(10\ 000 - 7\ 500)}{227(17\ 500 - 7\ 500)} = 0.517\ 62, \\ f_{227}(10\ 000) &= \frac{42}{227(17\ 500 - 7\ 500)} = 0.000\ 018\ 502.\end{aligned}$$

方差估计为

$$\begin{aligned}\widehat{\mathrm{Var}}[S_{227}(10\ 000)] &= \frac{1}{227(10\ 000)^2}\left[10\ 000^2\frac{99}{227}\frac{128}{227} + 2\ 500^2\frac{42}{227}\frac{185}{227}\right. \\ &\quad \left.-2(10\ 000)(2\ 500)\frac{99}{227}\frac{42}{227}\right] \\ &= 0.000\ 947\ 13,\end{aligned}$$

且

$$\widehat{\mathrm{Var}}[f_{227}(10\ 000)] = \frac{\frac{42}{227}\frac{185}{227}}{227(10\ 000)^2} = 6.642\ 7 \times 0^{-12}. \qquad \square$$

类似于数据集 A 中的离散数据可以看作分组数据的一种特殊形式, 每一个离散点的概率可看作某个区间的概率.

例 11.10 证明对于离散随机变量, 样本点的概率的经验估计是无偏且相合的, 并推导其方差的表达式.

证 令 N_j 为样本中 x_j 被观测到的次数, 则 N_j 服从参数为 n 和 $p(x_j)$ 二项分布. 经验估计的公式为 $p_n(x_j) = N_j/n$ 且有

$$\mathrm{E}[p_n(x_j)] = \mathrm{E}\left(\frac{N_j}{n}\right) = \frac{np(x_j)}{n} = p(x_j),$$

说明此估计值是无偏的. 另有

$$\mathrm{Var}[p_n(x_j)] = \mathrm{Var}\left(\frac{N_j}{n}\right) = \frac{np(x_j)[1-p(x_j)]}{n^2} = \frac{p(x_j)[1-p(x_j)]}{n},$$

当 $n \to \infty$ 时其值趋于零, 因此这个估计是相合的. □

例 11.11 基于数据集 A 计算 $p(2)$ 的经验估计并计算该估计量的方差.

解 经验估计为

$$p_{94\ 935}(2) = \frac{1\ 618}{94\ 935} = 0.017\ 043,$$

其方差的估计值为

$$\frac{0.017\ 043(0.982\ 957)}{94\ 935} = 1.764\ 66 \times 10^{-7}.$$ □

方差可用来构造未知概率的置信区间.

例 11.12 基于数据集 A, 用 (9.3) 式和 (9.4) 式构造 $p(2)$ 的近似 95%置信区间.

解 由 (9.3) 式知

$$0.95 = \Pr\left(-1.96 \leqslant \frac{p_n(2) - p(2)}{\sqrt{p(2)[1-p(2)]/n}} \leqslant 1.96\right).$$

为解出置信区间的端点值, 将不等式变为等式, 再两边平方后得到 (为了简便起见, 将原式中的 $p(2)$ 省略为 p)

$$\begin{aligned}\frac{(p_n - p)^2 n}{p(1-p)} &= 1.96^2,\\ np_n^2 - 2npp_n + np^2 &= 1.96^2 p - 1.96^2 p^2,\\ 0 &= (n + 1.96^2)p^2 - (2np_n + 1.96^2)p + np_n^2.\end{aligned}$$

进而解得

$$p = \frac{2np_n + 1.96^2 \pm \sqrt{(2np_n + 1.96^2)^2 - 4(n + 1.96^2)np_n^2}}{2(n + 1.96^2)},$$

这个式子给出了置信区间的两个端点值. 将数据集 A 中的数值 ($p_n = 0.017\,043, n = 94\,935$) 代入其中就可得到置信区间 (0.016 239, 0.017 886).

方程 (9.4) 直接给出了置信区间的端点值

$$p_n \pm 1.96\sqrt{\frac{p_n(1-p_n)}{n}}.$$

将数据集 A 中的数值代入其中就得到区间的端点值 $0.017\,043 \pm 0.000\,823$, 于是置信区间为 (0.016 220, 0.017 866). 两种方法得到的答案非常接近, 在大样本的前提下, 这正是我们希望的结果. 因为正态分布是二项分布的一个良好近似, 因此这个结果是十分合理的. □

当数据存在删失或截断的情况时, 问题将变得更为复杂. 计数变量不再服从二项分布, 导致估计量的分布更难得到. 以下得到的每个结果都可以被严格证明, 但这里并不提供这些证明过程, 而是尽力说明这些结果的合理性.

考虑 $S(t)$ 的 Kaplan-Meier 有限乘积估计. 它是一系列具有 $(r_j - s_j)/r_j$ 形式的项之积, 其中 r_j 看作是在年龄 y_j 面临身故风险的个体总数, 而 s_j 是实际身故的个体总数. 假设身故年龄以及在该年龄面临身故风险的个体数目是固定的, 则 s_j 是唯一的随机量. 作为随机变量 s_j 服从二项分布, 相应的 "试验总次数" 为 r_j, "成功" 概率为 $[S(y_{j-1}) - S(y_j)]/S(y_{j-1})$, 之所以是这样一个概率值, 是因为面临身故风险的个体在前一个身故年龄必然是存活的. 对于某一个上述的乘积项, 有

$$\mathrm{E}\left(\frac{r_j - S_j}{r_j}\right) = \frac{r_j - r_j[S(y_{j-1}) - S(y_j)]/S(y_{j-1})}{r_j} = \frac{S(y_j)}{S(y_{j-1})}.$$

即这个比值是对从某个死亡年龄生存到下一个死亡年龄的概率的无偏估计. 进而有

$$\begin{aligned}\mathrm{Var}\left(\frac{r_j - S_j}{r_j}\right) &= \frac{r_j \frac{S(y_{j-1})-S(y_j)}{S(y_{j-1})}\left[1 - \frac{S(y_{j-1})-S(y_j)}{S(y_{j-1})}\right]}{r_j^2} \\ &= \frac{[S(y_{j-1}) - S(y_j)]S(y_j)}{r_j S(y_{j-1})^2}.\end{aligned}$$

现在考虑一个指定的死亡年龄的生存概率的估计值, 其期望值为

$$\begin{aligned}\mathrm{E}[\hat{S}(y_j)] &= \mathrm{E}\left[\prod_{i=1}^{j}\left(\frac{r_i - S_i}{r_i}\right)\right] = \prod_{i=1}^{j}\mathrm{E}\left(\frac{r_i - S_i}{r_i}\right) \\ &= \prod_{i=1}^{j}\frac{S(y_i)}{S(y_{i-1})} = \frac{S(y_j)}{S(y_0)},\end{aligned}$$

其中 y_0 是样本中最小的观测年龄. 为了能够将期望值的符号放进乘积号之中, 这里假设所有的 S 值是相互独立的. 此结果说明在已知死亡发生的年龄时, 其生存函数的估计是无偏的.

考虑方差的计算时, 首先需要了解一个关于独立随机变量乘积的方差的一般性结论. 设 $X_1, \cdots, X_n$ 是相互独立的随机变量, 且 $\mathrm{E}(X_j) = \mu_j$, $\mathrm{Var}(X_j) = \sigma_j^2$, 则有

$$\begin{aligned}\mathrm{Var}(X_1 \cdots X_n) &= \mathrm{E}(X_1^2 \cdots X_n^2) - E(X_1 \cdots X_n)^2 \\ &= \mathrm{E}(X_1^2) \cdots \mathrm{E}(X_n^2) - E(X_1)^2 \cdots \mathrm{E}(X_n)^2 \\ &= (\mu_1^2 + \sigma_1^2) \cdots (\mu_n^2 + \sigma_n^2) - \mu_1^2 \cdots \mu_n^2.\end{aligned}$$

对于有限乘积估计, 有

$$\begin{aligned}\mathrm{Var}[S_n(y_j)] &= \mathrm{Var}\left[\prod_{i=1}^{j}\left(\frac{r_i - S_i}{r_i}\right)\right] \\ &= \prod_{i=1}^{j}\left[\frac{S(y_i)^2}{S(y_{i-1})^2} + \frac{[S(y_{i-1}) - S(y_i)]S(y_i)}{r_i S(y_{i-1})^2}\right] - \frac{S(y_j)^2}{S(y_0)^2} \\ &= \prod_{i=1}^{j}\left[\frac{r_i S(y_i)^2 + [S(y_{i-1}) - S(y_i)]S(y_i)}{r_i S(y_{i-1})^2}\right] - \frac{S(y_j)^2}{S(y_0)^2} \\ &= \prod_{i=1}^{j}\left[\frac{S(y_i)^2}{S(y_{i-1})^2}\frac{r_i S(y_i) + [S(y_{i-1}) - S(y_i)]}{r_i S(y_i)}\right] - \frac{S(y_j)^2}{S(y_0)^2} \\ &= \frac{S(y_j)^2}{S(y_0)^2}\left\{\prod_{i=1}^{j}\left[1 + \frac{S(y_{i-1}) - S(y_i)}{r_i S(y_i)}\right] - 1\right\}.\end{aligned}$$

用这个公式进行计算显得太繁琐, 因此常用以下介绍的近似. 这种近似方法是基于这样一个事实: 对于任一列比较小的数值 $a_1, \cdots, a_n$, 乘积 $(1 + a_1) \cdots (1 + a_n)$ 近似等于 $1 + a_1 \cdots + a_n$, 这是因为被略去的项都是两个或者更多 a_i 的乘积, 如果这些 a_i 本来就很小, 这些乘积值就会更小以至于可以被忽略. 运用这个结论得到近似计算公式

$$\mathrm{Var}[S_n(y_j)] = \left[\frac{S(y_j)}{S(y_0)}\right]^2 \sum_{i=1}^{j} \frac{S(y_{i-1}) - S(y_i)}{r_i S(y_i)}.$$

由于生存函数已知的情况是不可能出现的, 因此需要将式中的生存函数替换为其估计值. 回忆 $S(y_j)$ 的估计值其实是已知年龄 y_0 仍然存活的条件下的估计值, 并且 $(r_i - s_i)/r_i$ 是 $S(y_i)/S(y_{i-1})$ 的估计值, 因此有

$$\widehat{\mathrm{Var}}[S_n(y_j)] = S_n(y_j)^2 \sum_{i=1}^{j} \frac{s_i}{r_i(r_i - s_i)}. \tag{11.3}$$

方程 (11.3) 称为 Greenwood 近似公式 (Greenwood approximation), 上式为本书唯一的 Greenwood 近似公式.

例 11.13 使用数据集 D1 估计 $S_{30}(3)$, 先直接估计再使用 Greenwood 公式估计. 对 ${}_2\hat{q}_3$ 也进行同样的计算.

解　由于不存在删失和截断, 可以直接用经验公式来估计方差. 在 30 人中共有 3 人身故, 因此

$$\widehat{\text{Var}}[S_30(3)] = \frac{(3/30)(27/30)}{30} = \frac{81}{30^3}.$$

对于 Greenwood 近似, $r_1 = 30, s_1 = 1$, $r_2 = 29, s_2 = 2$. 近似值为

$$\left(\frac{27}{30}\right)^2 \left(\frac{1}{30(29)} + \frac{2}{29(27)}\right) = \frac{81}{30^3}.$$

可以证明, 如果没有删失和截断的情况, 这两个公式一定会得到相同的答案. 回忆 Greenwood 公式推导过程可以发现, 它仅可以在出现身故的年龄计算方差. 对于并没有身故发生的年龄, 习惯上将 Greenwood 公式中的求和上限规定为不超过该年龄的有身故事件发生的最大年龄值.

下面考虑 ${}_2\hat{q}_3$ 的计算. 和例 11.6 类似可得方差的 (条件) 估计值

$$\widehat{\text{Var}}({}_2\widehat{q}_3) = \frac{(4/27)(23/27)}{27} = \frac{92}{27^3}.$$

运用 Greenwood 公式计算时, 首先应注意到实际上需要估计的量是

$${}_2q_3 = \frac{S(3) - S(5)}{S(3)} = 1 - \frac{S(5)}{S(3)}.$$

在经验估计的过程中, 所有计算必须在给定第 3 期有 27 人存活的条件下进行, 进而 ${}_2\hat{q}_3$ 的方差和 $\hat{S}(5)$ 的方差一样, 只用到第 3 期及其以后的信息进行估算. 从第 3 期开始, 共有 3 个时刻有身故事件发生：3.1, 4.0 和 4.8, 并且 $r_1 = 27, r_2 = 26, r_3 = 25, s_1 = 1, s_2 = 1$, $s_3 = 2$. 其 Greenwood 近似为

$$\left(\frac{23}{27}\right)^2 \left(\frac{1}{27(26)} + \frac{1}{26(25)} + \frac{2}{25(23)}\right) = \frac{92}{27^3}.$$ □

例 11.14　**重新考虑上例, 将数据换成数据集 D2 中的所有 40 个观测值, 以及其他由于删失和截断导致不完整的信息, 其他条件和问题不变.**

解　这个例题不能直接运用经验方法解决, 因为这里的样本容量并不明确 (由于截断和极端情况的存在, 随时间的推移不同个体不断进入和退出, 进而样本容量随着时间在不断变化). 从例 11.2 可知前 3 年相关取值为 $r_1 = 30, r_2 = 26, s_1 = 1, s_2 = 2$; 由例 11.3 可知 $S_{40}(3) = 0.892\ 3$. 此时 Greenwood 估计是

$$(0.892\ 3)^2 \left(\frac{1}{30(29)} + \frac{2}{26(24)}\right) = 0.003\ 467\ 1.$$

运用正态近似, 可以构造一个近似 95%置信区间, 其端点值如下

$$0.892\ 3 \pm 1.96\sqrt{0.003\ 467\ 1} = 0.892\ 3 \pm 0.115\ 4,$$

得到的区间是 (0.776 9, 1.007 7). 当样本容量较小时, 置信区间的端点取值可能小于 0 或大于 1.

下面考虑 ${}_2\hat{q}_3$ 的计算. 相关的量 (从第 3 期开始, 仍沿用前例中使用的脚标) 为 $r_3 = 26$, $r_4 = 26$, $r_5 = 23, r_6 = 21$, $s_3 = 1, s_4 = 2$, $s_5 = 1$ 和 $s_6 = 1$, 方差的估计值为

$$\left(\frac{0.721\ 5}{0.892\ 3}\right)^2\left(\frac{1}{26(25)}+\frac{2}{26(24)}+\frac{1}{23(22)}+\frac{1}{21(22)}\right)=0.005\ 950\ 2. \qquad \square.$$

从上个例题可以看出, 使用通常的方法构造置信区间可能会得到不合理的结果. 下面用另一种方法构造. 令 $Y = \ln[-\ln S_n(t)]$, 运用 delta 方法 (详见定理 12.17) 来估计 Y 的方差. 其中, 定义函数 $g(x) = \ln(-\ln x)$, 求导有

$$g'(x) = \frac{1}{-\ln x}\frac{-1}{x} = \frac{1}{x\ln x}.$$

根据 delta 方法, Y 的方差可由以下公式近似估计

$$\{g'[\mathrm{E}(S_n(t))]\}^2\mathrm{Var}[S_n(t)] = \frac{\mathrm{Var}[S_n(t)]}{[S_n(t)\ln S_n(t)]^2},$$

其中用到 $S_n(t)$ 是 $S(t)$ 的无偏估计这个重要事实. 进一步估计 $\theta = \ln[-\ln S(t)]$ 的 95%置信区间的端点为

$$\ln[-\ln S_n(t)] \pm 1.96\frac{\sqrt{\widehat{\mathrm{Var}}[S_n(t)]}}{S_n(t)\ln S_n(t)}.$$

因为 $S(t) = \exp(-\mathrm{e}^{\theta})$, 将上述置信区间的端点带入此表达式就可以得到 $S(t)$ 的置信区间.

区间上限为 (令 $\hat{v} = \widehat{\mathrm{Var}}[S_n(t)]$)

$$\begin{aligned}&\exp\{-\mathrm{e}^{\ln[-\ln S_n(t)]+1.96\sqrt{\hat{v}}/[S_n(t)\ln S_n(t)]}\}\\ =&\exp\{[\ln S_n(t)]\mathrm{e}^{1.96\sqrt{\hat{v}}/[S_n(t)\ln S_n(t)]}\}\\ =&S_n(t)^U, \quad U = \exp\left[\frac{1.96\sqrt{\hat{v}}}{S_n(t)\ln S_n(t)}\right]\end{aligned}$$

类似地, 区间下界是 $S_n(t)^{1/U}$. 这个区间一定在 0 到 1 的区间之内, 通常称作对数转换的置信区间 (log-transformed confidence interval).

例 11.15 用对数转换的方法计算例 11.14 中 $S(3)$ 的置信区间.

解 由于

$$U = \exp\left[\frac{1.96\sqrt{0.003\ 467\ 1}}{0.892\ 3\ln(0.892\ 3)}\right] = 0.321\ 42.$$

因此区间的下界是 $0.892\,3^{1/0.321\,42} = 0.701\,50$, 上界是 $0.892\,3^{0.321\,42} = 0.964\,04$. □

对于 Nelson-Åalen 估计方法也可得到类似的结果. 以下介绍一个推导方差估计的直观方法. 在推导 Kaplan-Meier 估计量时, 所有结果都是在风险集已知且为非随机的确定量的假设下得到的, 以下的推导也沿用这样的假设. 在身故时刻 t_i 身故的个体数近似服从参数为 $r_i h(t_i)$ 的 Poisson 分布[①], 因此其方差是 $r_i h(t_i)$, 而这个方差的值可以由 $r_i(s_i/r_i) = s_i$ 来近似. 此时有 (仍然假设独立)

$$\widehat{\mathrm{Var}}[\hat{H}(y_j)] = \widehat{\mathrm{Var}}\left(\sum_{i=1}^{j} \frac{s_i}{r_i}\right) = \sum_{i=1}^{j} \frac{\widehat{\mathrm{Var}(s_i)}}{r_i^2} \doteq \sum_{i=1}^{j} \frac{s_i}{r_i^2}.$$

线性置信区间是

$$\hat{H}(t) \pm z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}[\hat{H}(y_j)]}.$$

和生存函数[②]置信区间的讨论类似, 对数转换的置信区间是

$$\hat{H}(t)U, \quad \text{其中 } U = \exp\left[\pm\frac{z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}[\hat{H}(y_j)]}}{\hat{H}(t)}\right].$$

例 11.16 利用上述各个公式构造 $H(3)$ 的近似 95%置信区间. 使用数据集 D2 中所有 40 个观测值.

解 点估计为 $\hat{H}(3) = \frac{1}{30} + \frac{2}{26} = 0.110\,26$, 方差的估计值为 $\frac{1}{30^2} + \frac{2}{26^2} = 0.004\,069\,7$, 线性置信区间的端点为

$$0.110\,26 \pm 1.96\sqrt{0.004\,069\,7} = 0.110\,26 \pm 0.125\,04,$$

所得到的区间是 $(-0.014\,78, 0.235\,30)$. 对数变换后有

$$U = \exp\left[\pm\frac{1.96(0.004\,069\,7)^{1/2}}{0.110\,26}\right] = \exp(\pm 1.134\,02) = 0.321\,74 \text{ 到 } 3108\,13.$$

可得从 0.110 26(0.321 74)=0.035 48 到 0.110 26(3.108 13)=0.342 70 的置信区间. □

习题

11.10 利用数据集 D1 的所有信息计算 $q_j, j = 0, \cdots, 4$ 以及 ${}_5p_0$ 的经验估计, 这里考虑退保事件. 并对以上的每个估计量给出估计的方差, 并指出哪个方差估计值是条件估计值. 将 ${}_5q_0$ 看做在 5 年保险期内退保的概率.

① 也可将其假设为二项分布 (与 Kaplan-Meier 的推导方法类似); 类似地, 也可将 Poisson 假设应用于 Kaplan-Meier 估计量的推导. 这里给出的是最常用的公式.

② 该区间的推导用到了如下的变换: $Y = \ln \hat{H}(t)$.

11.11 对于数据集 A 用经验估计法计算出现 2 次以上事故的概率, 并计算这个估计值的方差.

11.12 在例 11.13 中考虑退保事件作为变量, 其他条件不变, 重新计算所有问题.

11.13 在例 11.14 中考虑退保事件, 其他条件不变, 重新计算所有问题. 将 ${}_2q_3$ 看作在 5 年保险期内退保的概率.

11.14 用例 11.13 的条件计算 $S(3)$ 的对数转换的置信区间.

11.15 考虑退保事件, 用数据集 D2 中的全部 40 个观测值构造 $H(3)$ 的 95%置信区间.

11.16* 现有 10 个从出生便开始观测直至其身故的个体, 所有个体的身故年龄如表 11-4 所示. 在不作任何有关分布类型的假设下计算的 ${}_3\hat{q}_7$, 其条件方差记为 V_1; 另一方面, 如果在已知生存函数是 $S(t) = 1 - t/15$ 的情况下估计 ${}_3\hat{q}_7$, 其条件方差记为 V_2. 求 $V_1 - V_2$.

11.17* 在 0 到 1 年的区间中, 面对死亡威胁的个体数 (r) 为 15, 身故个体数 (s) 为 3; 在 1 到 2 年的区间中, 面对死亡威胁的个体数和身故个体数分别为 80 和 24; 在 2 到 3 年的区间中, 这 2 个量分别为 25 和 5; 在 3 到 4 年的区间中, 这 2 个量变成 60 和 6; 在 4 到 5 年的区间中, 这 2 个量是 10 和 3. 用 Greenwood 近似公式计算 $\hat{S}(4)$ 的方差.

表 11-4　习题 11.16 的数据

年龄	身故人数
2	1
3	1
5	1
7	2
10	1
12	2
13	1
14	1

11.18* 假设观测值可能删失但没有截断发生, 并设 y_j, y_{j+1} 为 2 个连续的身故时刻. 用 Nelson-Åalen方法估计95%线性置信区间, $H(y_j)$的置信区间是(0.071 25, 0.228 75), $H(y_{j+1})$ 的置信区间是 (0.156 07, 0.386 35). 求 s_{j+1}.

11.19* 在一项死亡力研究中, 观测从出生开始的 50 个个体, 其中: 在年龄为 15 时有 2 个人身故, 年龄为 17 时有 3 个观测值删失, 年龄为 25 时有 4 个人身故, 年龄为 30 时有 c 个观测值删失, 年龄为 32 时有 8 个人身故, 年龄为 40 时有 2 个人身故. 用 S 表示 $S(35)$ 的有限乘积估计值, V 表示用 Greenwood 方法计算的 S 的方差. 已知 $V/S^2=$ 0.011 476, 求 c.

11.20* 从确诊时刻开始观测 15 例癌症患者, 直至其身故或者 36 个月后观测期结束, 结果如下: 15 个月时有 2 人身故, 20 个月时有 3 人身故, 24 个月时有 2 人身故, 30 个月时

有 d 人身故, 34 个月时有 2 人身故, 36 个月时有 1 人身故. $H(35)$ 的 Nelson-Åalen 估计值是 1.564 1, 求这个估计量的方差.

11.21* 已知数据如表 11-5 所示, 计算在时刻 20 的累积风险率函数的 Nelson-Åalen 估计量的标准差.

表 11-5　习题 11.21 的数据

y_j	r_j	s_j
1	100	15
8	65	20
17	40	13
25	31	31

11.3　核密度模型

经验分布的一个问题是它的离散性. 如果已知真实的分布是连续的, 则经验分布的近似程度会显得很差. 本节将介绍一种和经验分布相似的光滑的分布函数的估计方法. 回忆定义 10.4 的内容, 可以用连续的随机变量来替代每一个离散的概率, 习惯上要求连续随机变量的均值等于它所替代的数据点, 但这并不是必须的, 而仅仅是为了让核估计和经验估计有相同的均值. 对于这种模型, 一种思路是, 最终的观测值是分两步得到的: 第一步是从经验分布中随机地取得一个值; 第二步是从一个连续型的分布函数中随机地取得一个值, 这个连续型分布的均值恰好是第一步取到的数值. 此处提到的连续型分布称为核 (kernel) 函数.

首先引入一些记号: 用 $p(y_i)$ 表示在经验分布的假设下 y_i 的概率 $(j=1,\cdots,k)$, $K_y(x)$ 表示均值为 y 的连续型随机变量的分布函数, 用 $k_y(x)$ 表示与该分布函数对应的密度函数.

定义 11.17　*分布函数的核密度估计*(kernel density estimator) *定义为*

$$\hat{F}(x)=\sum_{j=1}^{k}p(y_j)K_{y_j}(x).$$

密度函数的估计是

$$\hat{f}(x)=\sum_{j=1}^{k}p(y_j)k_{y_j}(x).$$

函数 $k_y(x)$ 称为核函数, 常见的有以下 3 种核函数.

定义 11.18 均匀核函数(uniform kernel) 定义为

$$k_y(x)=\begin{cases}0, & x<y-b,\\ \dfrac{1}{2b}, & y-b\leqslant x\leqslant y+b,\\ 0, & x>y+b,\end{cases}$$

$$K_y(x)=\begin{cases}0, & x<y-b,\\ \dfrac{x-y+b}{2b}, & y-b\leqslant x\leqslant y+b,\\ 1, & x>y+b.\end{cases}$$

三角核函数(triangular kernel) 定义为

$$k_y(x)=\begin{cases}0, & x<y-b,\\ \dfrac{x-y+b}{b^2}, & y-b\leqslant x\leqslant y,\\ \dfrac{y+b-x}{b^2}, & y\leqslant x\leqslant y+b,\\ 0, & x>y+b,\end{cases}$$

$$K_y(x)=\begin{cases}0, & x<y-b,\\ \dfrac{(x-y+b)^2}{2b^2}, & y-b\leqslant x\leqslant y,\\ 1-\dfrac{(y+b-x)^2}{2b^2}, & y\leqslant x\leqslant y+b,\\ 1, & x>y+b.\end{cases}$$

令核函数是形状参数为 α, 尺度参数为 y/α 的 gamma 分布的密度函数, 就得到了**gamma核函数**(Gamma kernel), 具体表达如下:

$$K_y(x)=\frac{x^{\alpha-1}\mathrm{e}^{-x\alpha/y}}{(y/\alpha)^{\alpha}\Gamma(\alpha)}.$$

注意到此处 gamma 分布的均值为 $\alpha(y/\alpha)=y$, 方差为 $\alpha(y/\alpha)^2=y^2/\alpha$.

在每个核函数的定义中, 都有一个参数控制着核函数的分散度: 在前两个定义中, 这个参数是 $b(b>0)$, b 也称作带宽 (bandwidth); 在 gamma 核函数中参数 α 控制函数的分散度, α 越大函数的分散度越小. 也有一些核函数在零到无穷的范围内都有非零取值.

例 11.19 利用核密度估计方法以及上述 3 种核函数, 计算例 10.8 中的估计量.

解 由经验分布模型, 取 1.0 的概率是 1/8, 取 1.3 的概率是 1/8, 取 1.5 的概率是 2/8, 取 2.1 的概率是 3/8, 取 2.8 的概率是 1/8. 若采用带宽为 0.1 的均匀核函数, 效果并不明显. 数据点 1.0 由 0.9 到 1.1 之间的水平密度函数表示, 其高度是 (1/8)[1/2(0.1)] = 0.625; 另外, 若带宽为 1.0, 数据点 1.0 由 0.0 到 2.0 之间高度是

$(1/8)[1/2(1)]=0.0625$ 的水平密度函数表示. 图 11-1 和 11-2 描出了上述密度函数的图形.

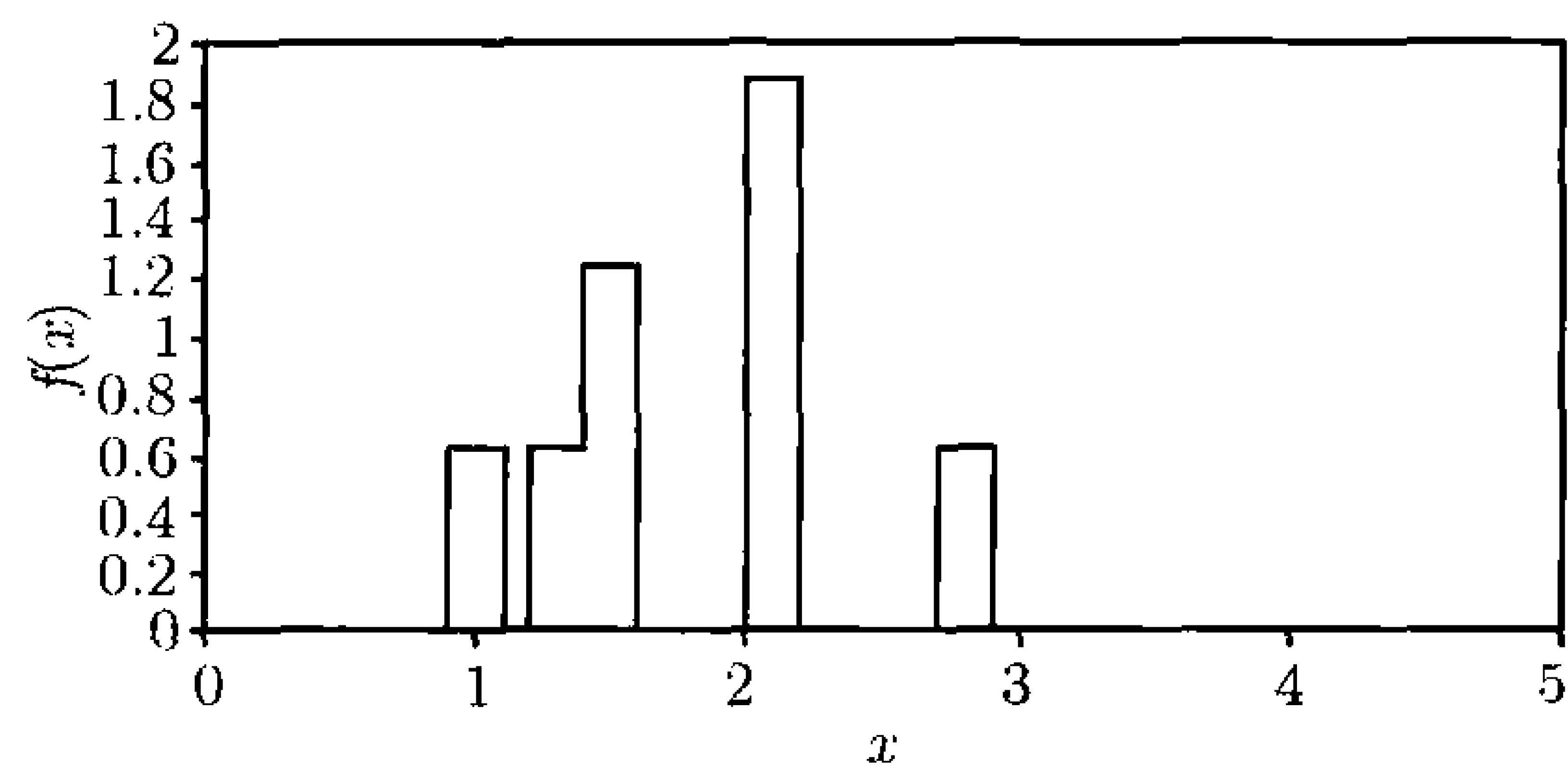


图 11-1　带宽为 0.1 的均匀核密度

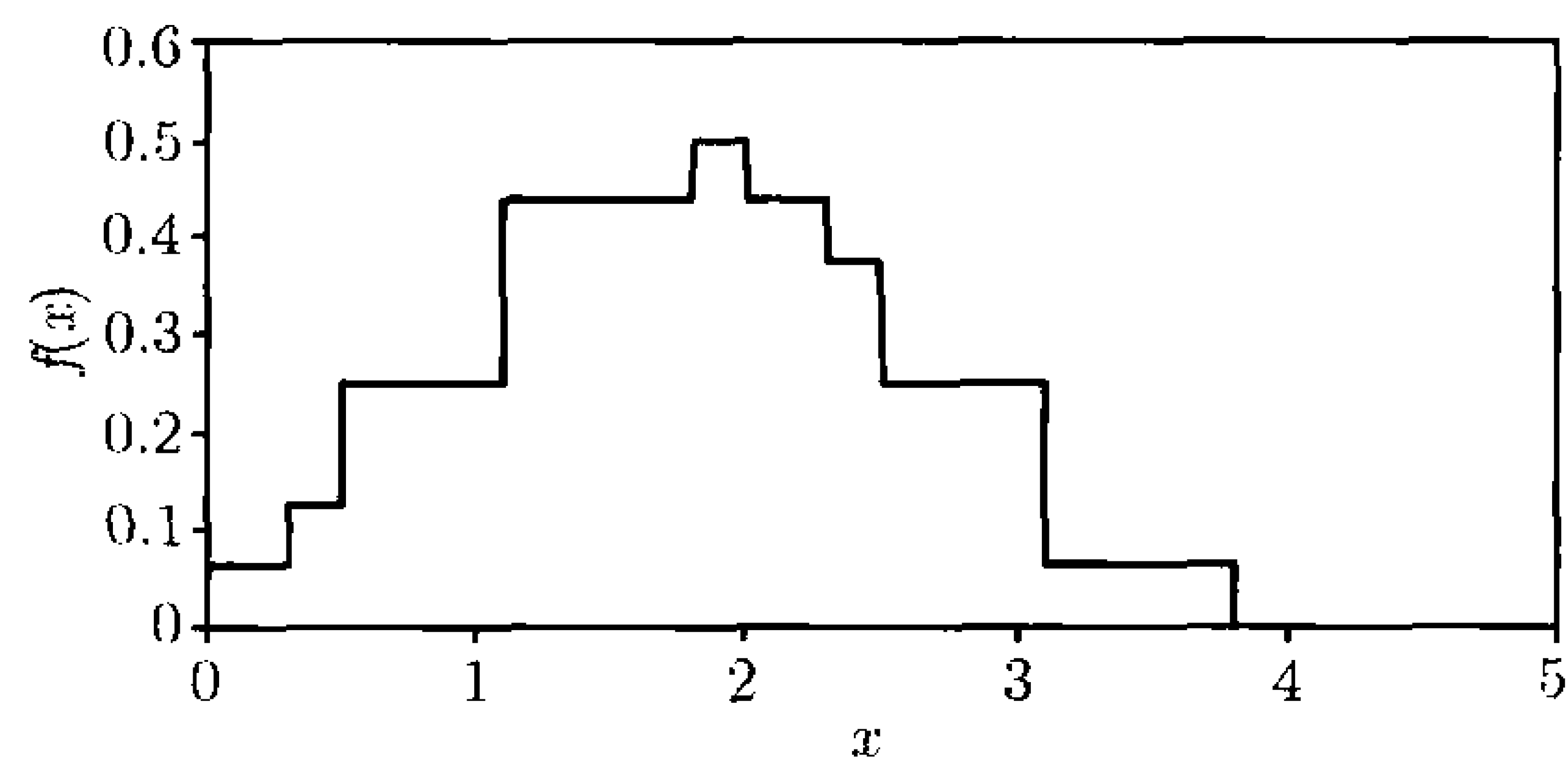


图 11-2　带宽为 1.0 的均匀核密度

需要看清的一点是, 更大的带宽能够得到更光滑的效果. 从极限的观点看, 如果带宽趋于零, 核密度的估计方法就和经验估计一样了. 但也要注意如果带宽过大, 则可能在负值上出现概率, 产生不合理的结果. 有许多方法可以解决这个问题, 但这里不作介绍了.

采用三角核函数, 则每个点被一个三角形取代, 图 11-3 和图 11-4 显示了三角核函数在前文讨论过的两个带宽下的图形.

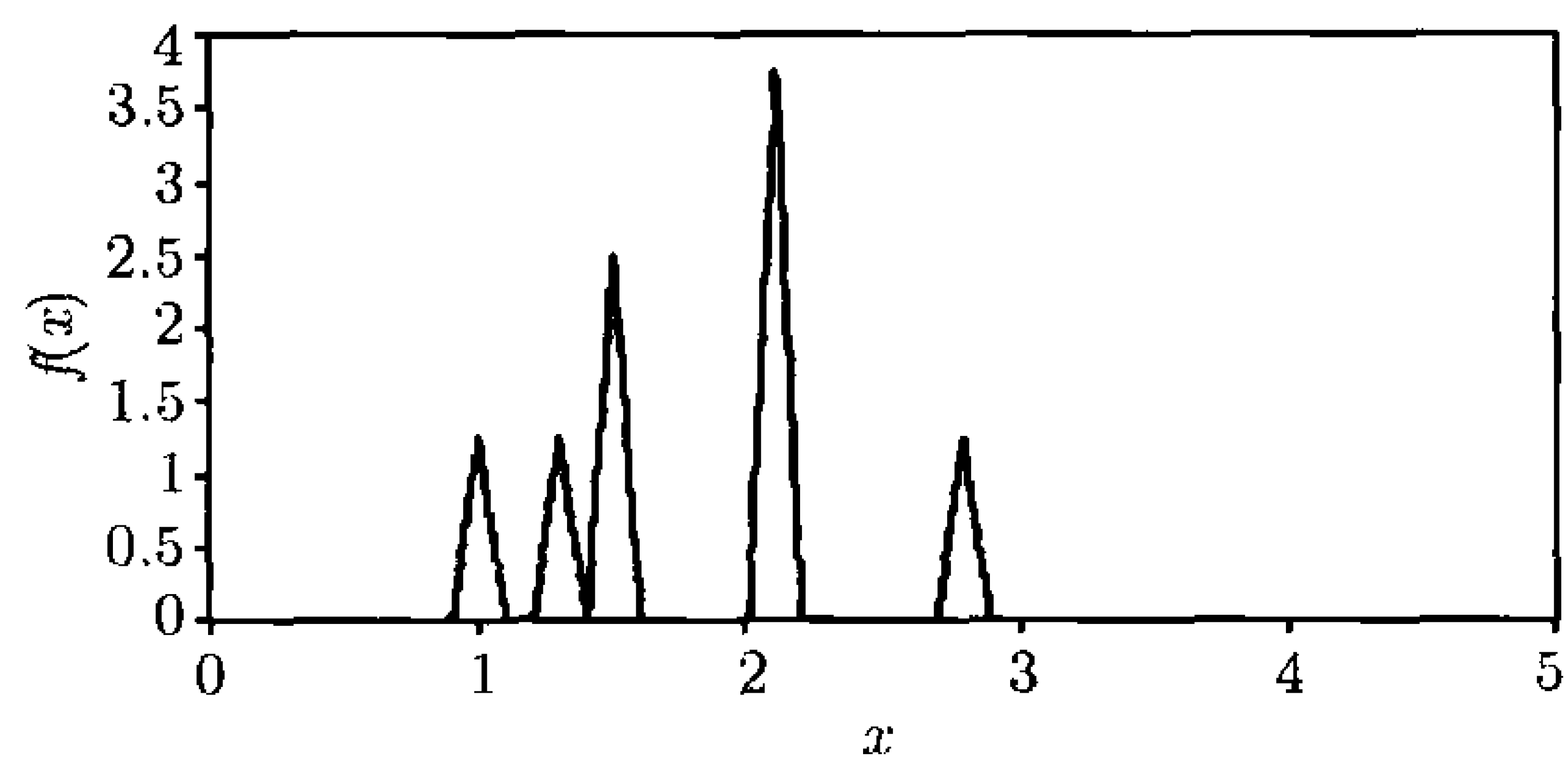


图 11-3　带宽为 0.1 的三角核密度

同样, 更大的带宽能够得到更光滑的效果. gamma 核函数实质上是一个混合

gamma 分布, 其中各数据点为每个 gamma 分布的均值, 经验分布的概率成为一个权重值. 该分布的密度函数是

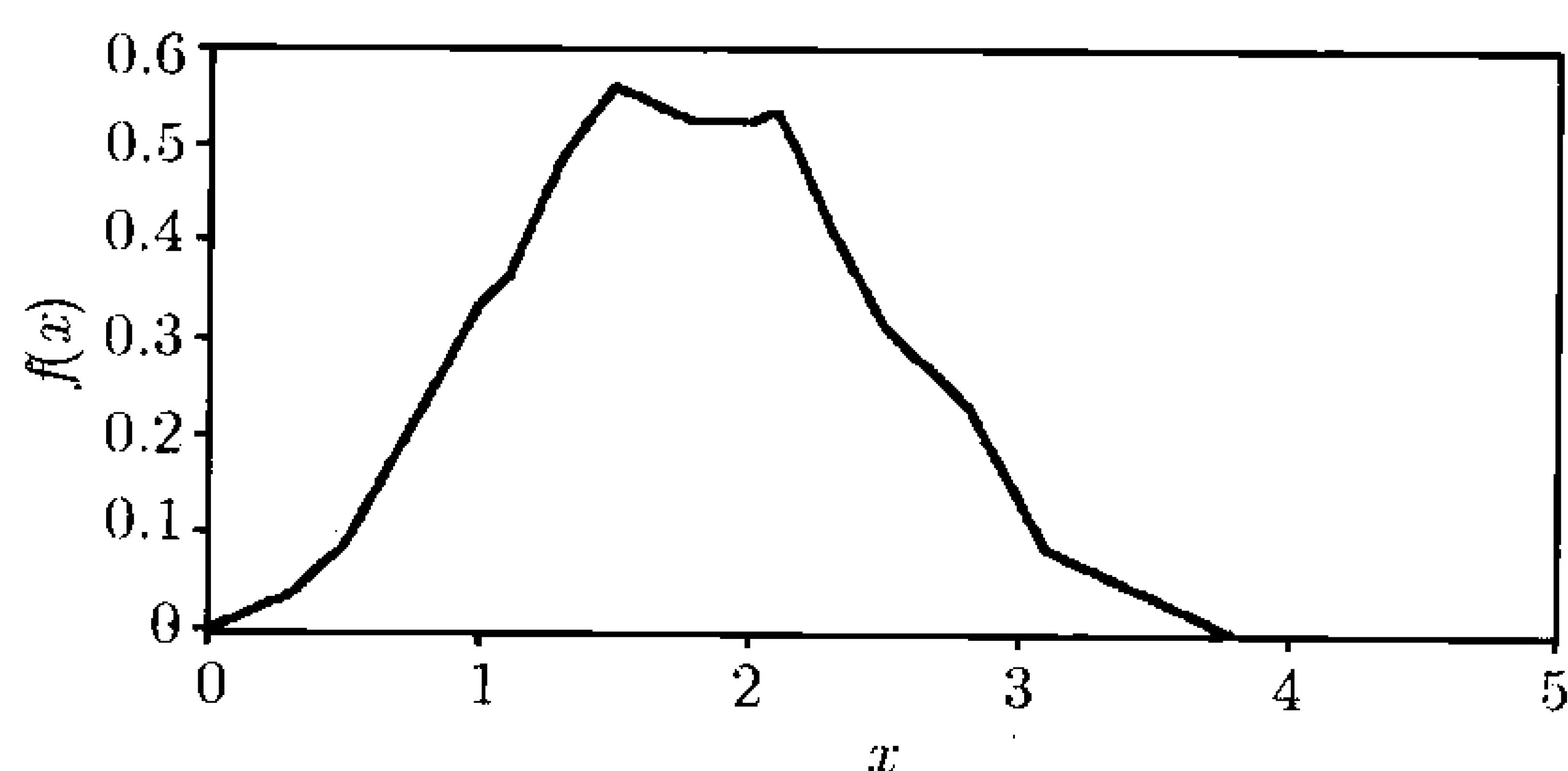


图 11-4 带宽为 1.0 的三角核密度

$$f_\alpha(x)=\sum_{j=1}^{5}p(y_j)\frac{x^{\alpha-1}\mathrm{e}^{-x\alpha/y_j}}{(y_j/\alpha)^\alpha\Gamma(\alpha)},$$

它的图形如图 11-5 和图 11-6 所示, α 分别取 500 和 50①. 对于这个核函数, 减小 α

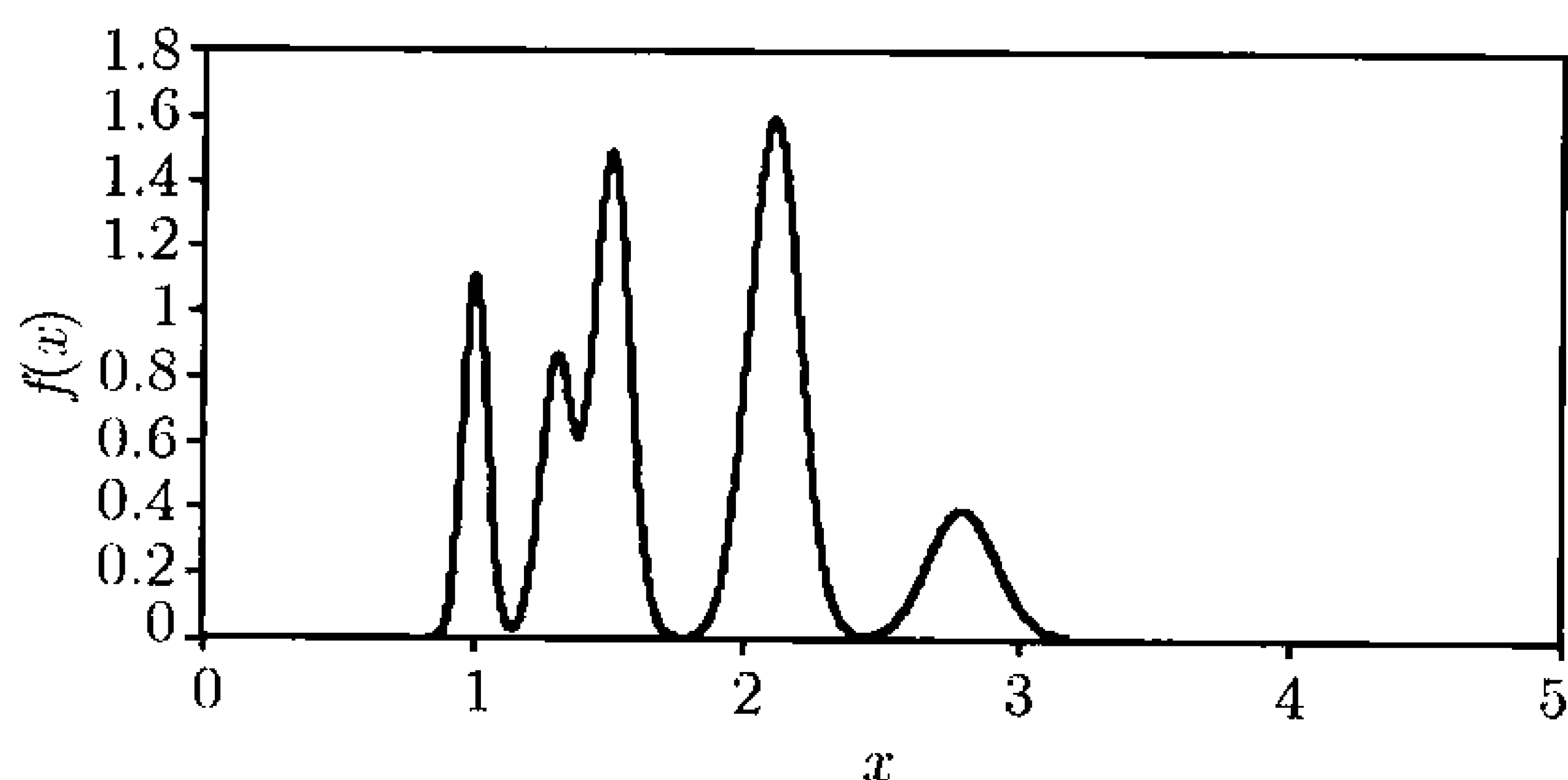


图 11-5 gamma 核密度, $\alpha=500$

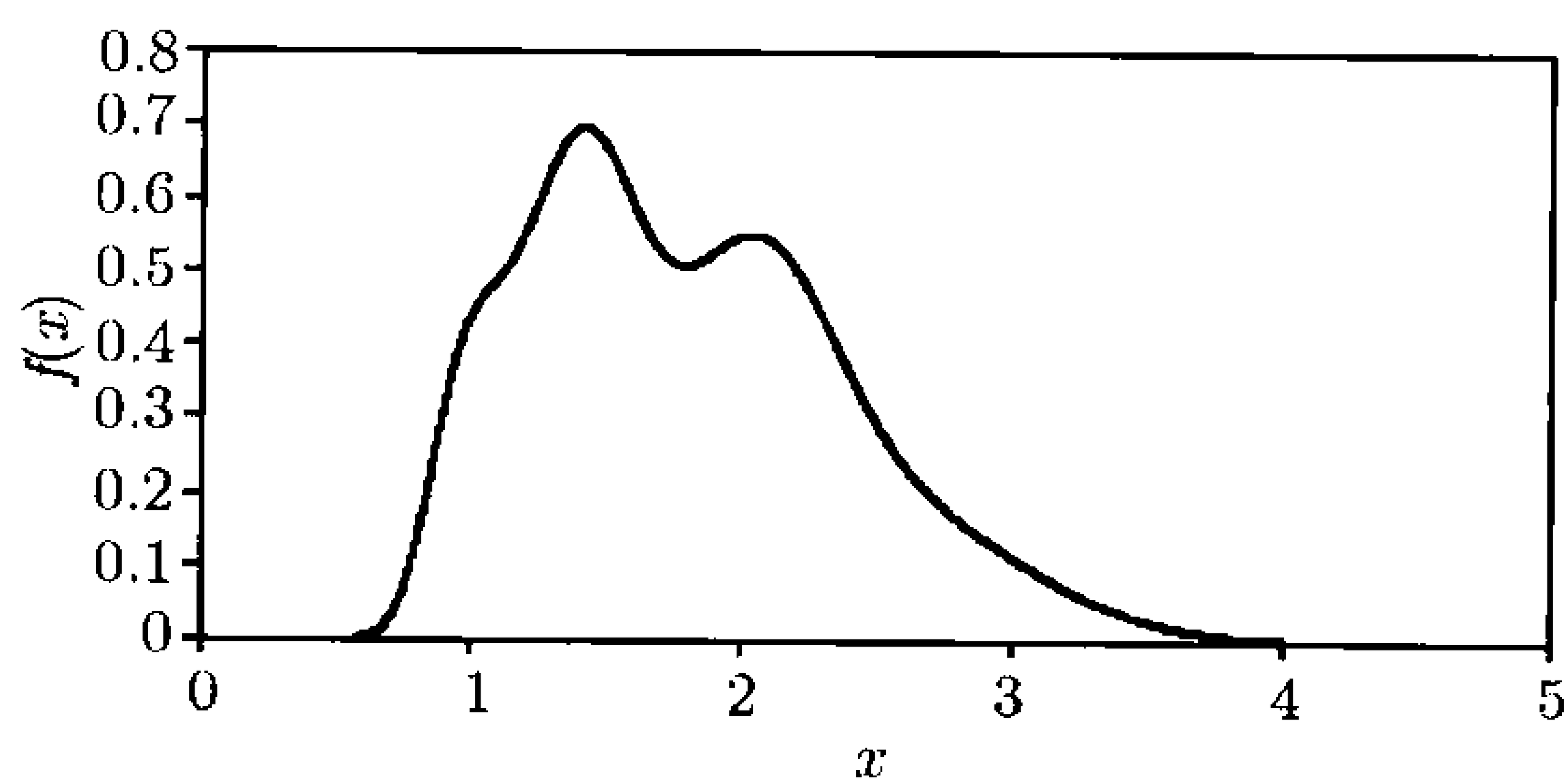


图 11-6 gamma 核密度, $\alpha=50$

① 在计算密度函数的时候, 可以采用对元素的比值求对数的方法减少上下溢出的问题, 即 $(\alpha-1)\ln x-x\alpha/y_j-\alpha\ln(y_j/\alpha)-\ln\Gamma(\alpha)$, 然后对所得结果取幂.

的值使得光滑程度变得更高. 文献 [23] 有关于 gamma 核函数的进一步讨论, 作者认为较优的 α 值是 $\sqrt{n}/(\hat{\mu}_4'/\hat{\mu}_2'-1)^{1/2}$. □

习题

11.22 给出 Pareto 核函数的计算公式.

表 11-6 习题 11.24 的数据

t_j	s_j	r_j
10	1	20
34	1	19
47	1	18
75	1	17
156	1	16
171	1	15

11.23 用数据集 D2 构造退保时间的核密度估计. 要注意此处需要估计的分布是一个混合型分布 (概率函数在 0 到 5 是连续的, 但在 5 是间断的).

11.24* 死亡时间数据如表 11-6 所示, 用带宽为 60 的均匀核函数计算 $\hat{f}(100)$.

11.4 大数据集合的近似计算

11.4.1 引言

当数据量很大的时候, 如果使用 Kaplan-Meier 估计方法, 则必须进行大量的数据排序和计算工作, 而从结果看, 如此繁杂的工作显得冗余, 尤其是有时我们并不需要得到分布函数在所有点的值. 例如, 在构造生命表时, 只需要计算整数年龄的概念函数值即可. 关于生命表构造的更多细节和方法, 可以参阅 Batten[11] 和 London[85]. 虽然, 以下的讨论都是针对生命表构造进行的, 但这些方法也适用于任何其他数据的处理.

给定区间的端点 $c_0<c_1<\cdots<c_k$, 令 d_j 是在区间 $[c_j,c_{j+1})$ 中的某个点左截断的观测值的总数. 在进行生命表研究中, 这就是在给定范围内的某个年龄初次观测的个体数目. 类似地, 令 u_j 是在区间 $(c_j,c_{j+1}]$ 中的某个点右删失的观测值的总数. 请注意以上两种区间在端点处的区别. 之所以这样表示, 是因为截断可能发生在第一个区间的左端点但不可能发生在最后一个区间的右端点. 而对于删失, 情况正好相反. 应注意到, 所有观测值都和某个 d_j 相对应, 但只有出现删失的数据才会和某个 u_j 相对应. 用 x_j 表示区间 $(c_j,c_{j+1}]$ 中未删失的观测值数目, 则 $n=\sum\limits_{j=0}^{k-1}d_j=\sum\limits_{j=0}^{k-1}(u_j+x_j)$, 其中 n 是样本容量. 这种处理在计算中的优点体现在, 可以对数据集合中的数值逐步累加, 相当于需要处理的数据集合被减少了.

11.4.2 Kaplan-Meier 近似

为了运用 Kaplan-Meier 公式, 必须对每个区间内数值点的位置作一些假设. 最简单的假设是规定区间中所有未删失的数据都等于同一个值, 记这个值为 c_j^*, 所有左截断的值小于 c_j^*, 而所有右删失的值大于或等于 c_j^*. 则在 c_j^* 处的风险集是 $r_j = \sum_{i=0}^{j} d_i - \sum_{i=0}^{j-1}(u_i + x_i)$. 接下来的处理, 习惯上先估计分布函数在给定端点处的值, 再用插值方法 (通常是线性插值) 得到相邻两个值之间的光滑函数, 而不是在 c_j^* 点分配所有的概率 (正如 Kaplan-Meier 方法那样). 此时有

$$\begin{aligned}\hat{F}(c_0) &= 0,\\ \hat{F}(c_j) &= 1 - \prod_{i=0}^{j-1}\left(1 - \frac{x_i}{r_i}\right), \quad j = 1, 2, \cdots, k.\end{aligned} \tag{11.4}$$

令

$$q_j = {}_{c_{j+1}-c_j}q_{c_j} = \frac{S(c_j) - S(c_{j+1})}{S(c_j)}.$$

由 (11.4) 式得

$$\hat{q}_j = \frac{\prod_{i=0}^{j-1}\left(1 - \frac{x_i}{r_i}\right) - \prod_{i=0}^{j}\left(1 - \frac{x_i}{r_i}\right)}{\prod_{i=0}^{j-1}\left(1 - \frac{x_i}{r_i}\right)} = 1 - \left(1 - \frac{x_j}{r_j}\right) = \frac{x_j}{r_j}. \tag{11.5}$$

这是传统的生命表估计, 其中分子是已观测到的身故人数, 分母是面对风险的个体数 (即面临身故风险的人数). 在此公式中, 所有在当前时段或者更早的时刻进入研究项目的个体都认为是可能身故的个体, 但所有在当前时段之前就离开研究项目的个体都没有考虑在内. 如果需要研究的是货币额, 并且将所有可能的免赔额和限额都作为区间的端点, 上述公式恰好是在给定值处的有限乘积估计. 在生命表研究中, 这相当于规定所有个体都只能在其生日时加入研究项目 (按保险年龄看这是有可能的, 见习题 11.26) 或退保.

(11.5) 式可作以下推广. 令 $P_j = \sum_{i=0}^{j-1}(d_i - u_i - x_i)$ 为 c_j 年龄处于被观测状态的人数. 进一步假设所有的未删失的观测值都发生在时段中某个固定的点, 而且到这个时刻, 已经有 $100\alpha\%$ 的人进入研究项目 (计入 d_j), 总删失人数中有 $100\beta\%$ 的人已经删失 (计入 u_j), 此时的风险集为 $r_j = P_j + \alpha d_j - \beta u_j$. 公式 (11.5) 实际上为 $\alpha = 1, \beta = 0$ 时的特殊情况. α 和 β 的另一种选择是令 $\alpha = \beta = 0.5$, 相当于假设进入和退出研究项目的事件在区间内均匀分布, 此时有 $r_j = P_j + 0.5(d_j - u_j)$. 由 $P_{j+1} = P_j + d_j - u_j - x_j$, r_j 的表达式也可以写为

$$r_j = 0.5(P_j + P_{j+1} + x_j), \tag{11.6}$$

这个公式在生命表构造中十分常用. 有时也可能需要对不同的区间作不同的假设, 见例 11.20.

11.4.3　多元衰减表

下文中所有的估计工作的目的都是为了推导在没有截断和删失的情况下指定变量的概率分布. 对于损失额数据, 不考虑含免赔和限额的情况; 对于生存时间数据, 只考虑能够对每个个体从出生到身故跟踪观测的情况. 在文献 [16] 中, 将某个因素单独引起的 "死亡" 概率称作这个因素的**单衰减因子**(single-decrement rate), 记为 q_j'. 实际工作中常需要考虑**多元衰减概率**(multiple-decrement probability), 用 q_j 表示, 并将 q_j 制成表格. 用上标表明衰减原因, 例如, 假设衰减是由身故 (d)、退保 (w) 和退休 (r) 造成的, 则 $q_j^{'(w)}$ 表示在身故和退休都没有发生的前提下, 年龄为 c_j 的个体在 c_{j+1} 年龄之前退保的概率; 在身故和退保可能发生从而使得不可能由退休造成衰减的情况下, 年龄为 c_j 的个体在 c_{j+1} 年龄之前退休的概率记为 $q_j^{(r)}$. 多元衰减表的构造常常是先通过各种途径获得单个因素的衰减因子, 再用以下公式将这些因子共同影响下的概率表示出来:

$$q_j^{(g)} = \frac{\ln(1-q_j^{'(g)})}{\ln(1-q_j^{(T)})} q_j(T), q_j(T) = 1 - \prod_{g=1}^{m}(1-q_j^{'(g)}), \tag{11.7}$$

其中 g 为某个衰减因素的代表, m 是衰减因素的总数.

例 11.20　*运用本节介绍的方法以及数据集 D2 中的数据, 根据合理假设, 估计单个衰减因子以及多元衰减的相关量.*

解　首先考虑身故引起的衰减. 使用本节的记号, 相关的量如表 11-7 所示. 为了得到表中的结果, 一些假设是必要的. 考虑 $d_0 = 32$, 容易看出有 30 个等于 0 的值 (从一开始就跟踪观测的保单), 但对另外 2 份在启动之后才加入的保单, 要做出必要的假设. 在这里假设均匀分布即可, 即 $r_0 = 30 + 0.5(2) - 0.5(3) = 29.5$, 其余 r 值由 (11.6) 式计算, 请注意假设 33 号保单是在中点, 即 1.5 时刻加入的. 另需注意有

表 11-7　例 11.20 中单个衰减因素的计算

j	d_j	u_j	x_j	P_j	r_j	$q_j^{'(d)}$	$\hat{F}(j)$
0	32	3	1	0	29.5	0.033 9	0.000 0
1	2	2	0	28	28.0	0.000 0	0.033 9
2	3	3	2	28	28.0	0.071 4	0.033 9
3	3	3	3	26	26.0	0.115 4	0.102 9
4	0	21	2	23	21.0	0.095 2	0.206 4
5							0.282 0

17 份在 5 年后仍然有效的保单, 并假设这些保单都是在时刻 5 删失的, 而非均匀地分布于第 5 年中.

对于退保的情况, 利用公式 (11.7) 就可得到表 11-8 的多元衰减的相关数值 $q_j^{\prime(w)}$. □

表 11-8 多元衰减的计算

j	$q_j^{\prime(d)}$	$q_j^{\prime(w)}$	$q_j^{(T)}$	$q_j^{(d)}$	$q_j^{(w)}$
0	0.033 9	0.098 4	0.128 9	0.032 2	0.096 7
1	0.000 0	0.069 0	0.069 0	0.000 0	0.069 0
2	0.071 4	0.105 3	0.169 2	0.067 6	0.101 5
3	0.115 4	0.115 4	0.217 5	0.108 7	0.108 7
4	0.095 2	0.181 8	0.259 7	0.086 4	0.173 3

例 11.21 *现有一批保单损失额数据, 其中存在 0, 250 和 500 三种免赔额以及 5 000, 7 500 和 10 000 几种保单限额, 如表 11-9 所示. 用本节介绍的方法估计损失额的分布函数.*

表 11-9 例 11.21 的数据

范围	免赔额			总数
	0	250	500	
0~100	15			15
100~250	16			16
250~500	34	96		130
500~1 000	73	175	251	499
1 000~2 500	131	339	478	948
2 500~5 000	83	213	311	607
5 000~7 500	12	48	88	148
7 500~10 000	1	4	11	16
恰好 5 000	7	17	18	42
恰好 7 500	5	10	15	30
恰好 10 000	2	1	4	7
总数	379	903	1 176	2 458

解 计算过程和结果如表 11-10 所示. 由于免赔额和保单限额都处在区间端点处, 因此唯一合理的假设是 $\alpha = 1, \beta = 0$. □

表 11-10　例 11.21 的计算

c_j	d_j	u_j	x_j	P_j	r_j	$q_j^{'(d)}$	$\hat{F}(c_j)$
0	379	0	15	0	379	0.039 6	0.000 0
100	0	0	16	364	364	0.044 0	0.039 6
250	903	0	130	348	1 251	0.103 9	0.081 8
500	1 176	0	499	1 121	2 297	0.217 2	0.177 2
1 000	0	0	948	1 798	1 798	0.527 3	0.356 0
2 500	0	42	607	850	850	0.714 1	0.695 5
5 000	0	30	148	201	201	0.736 3	0.913 0
7 500	0	7	16	23	23	0.695 7	0.977 0
10 000							0.993 0

习题

11.25　给出表 11-8 的计算过程和结果.

11.26　在寿险产品销售时, 习惯上将被保险人的年龄指定为一个整数, 然后按年龄对应保费. 这时, q_x 的含义不再是 “x 岁个体在下一年死亡的概率”, 而是 “在 k 年前保单出售时为 $x-k$ 年龄的某人在今后一年内身故的概率是多少? ” 这种方式确定的年龄称作 “保险年龄”. 这种方法能够将被保险人的生日设置为保单的出售日, 从而使被保险的个体从其生日 (指定为保单出售日) 开始处于观测状态, 进而 d 的值不再是估计值而是准确值. 但是, 退保可以在任何时刻发生, 因此令 $\beta=0.5$ 是一个合理的假设①. 对于表 11-11 的数据, 分别采用精确的 Kaplan-Meier 估计以及本节介绍的方法估计 q_{45} 和 q_{46}.

表 11-11　习题 11.26 的数据

d	u	x	d	u	x
45	46.0		45	45.8	
45	46.0		46	47.0	
45		45.3	46	47.0	
45		46.7	46	46.3	
45		45.4	46		46.2
45	47.0		46		46.4
45	45.4		46	46.9	

11.27　现有 22 个保险赔付额的记录, 见表 11-12. 请合理设置尽可能少的区间, 用本节介绍的方法估计免赔额为 500 的保单其赔付额超过 5 000 的概率.

① 不过, 正如例 11.20 中对不同的 d 值采用不同的计算方法一样, 这里也可能对不同的 u 值采用不同的计算. 个体数据的观测可能因为个体的离开或者因为观测期限的截止而结束. 在研究个体时, 在保单出售后整数年时刻结束观测是很常见的情况, 这时保险年龄是整数, 对于这样的情况, $\beta=0$ 才是合理的假设.

表 11-12 习题 11.27 的数据

免赔额	赔付额	免赔额	赔付额
250	2 221	500	3 660
250	2 500*	500	215
250	207	500	1 302
250	3 735	500	10 000
250	5 000*	1 000	1 643
250	517	1 000	3 395
250	5 743	1 000	3 981
500	2 500*	1 000	3 836
500	525	1 000	5 000
500	4 393	1 000	1 850
500	5 000*	1 000	6 722

* 赔付额恰好是保单限额

第四部分　参数化统计方法

第12章　参数估计

在对某个现象采用参数模型建模时, 参数的确定是非常必要的. 可以采用任意的方法确定模型的参数, 但是基于对现象的观测数据来选择参数会更加合理. 特别地, 我们假设已经获得了 n 个独立的观测. 对于某些技术上的处理有时候进一步假设所有这些观测来自同一个随机变量, 而对于其他处理有时会放宽这个限制.

12.1 节中介绍的方法实现起来比较容易, 但得到的结果不太令人满意. 12.2 节阐述的最大似然估计方法在使用上比较困难, 但却具有很好的统计性质, 并且适用性更强.

12.1　矩方法和分位点匹配

在使用这些方法时假设 n 个观测都来自于同一个参数分布. 特别地, 设分布函数的形式如下

$$F(x/\boldsymbol{\theta}),\quad \boldsymbol{\theta}^{\mathrm{T}}=(\theta_1,\theta_2,\cdots,\theta_p),$$

其中 $\boldsymbol{\theta}^{\mathrm{T}}$ 表示 $\boldsymbol{\theta}$ 的转置, $\boldsymbol{\theta}$ 为列向量, 含有 p 个待估参数. 进一步, 令 $\mu_k'(\boldsymbol{\theta})=\mathrm{E}(X^k|\boldsymbol{\theta})$ 表示 k 阶原点矩, 令 $\pi_g(\boldsymbol{\theta})$ 表示随机变量的 $100g$ 分位点, 即: $F[\pi_g(\boldsymbol{\theta})|\boldsymbol{\theta}]=g$. 如果分布函数是连续的, 则这个等式至少有一个解.

对于来自同一个随机变量的 n 个独立观测样本, 令 $\hat{\mu}_k'=\frac{1}{n}\sum\limits_{j=1}^{n}x_j^k$ 为 k 阶矩的经验估计, 记 $\hat{\pi}_g$ 为 $100g$ 分位点的经验估计.

定义 12.1　*参数 $\boldsymbol{\theta}$ 的矩方法估计为下面 p 个方程的任意解*

$$\mu_k'(\boldsymbol{\theta})=\hat{\mu}_k',\quad k=1,2,\cdots,p.$$

这种估计方法的出发点是构造一个模型, 使其和实际数据有相同的前 p 阶矩 (实际数据的矩按照经验分布计算). 传统上矩估计的定义只对正整数矩, 这里也可以定义为负数矩或者分数矩. 特别地, 在考虑各种逆分布的参数估计时, 负数矩匹配可能是更好的方法①.

例 12.2　*使用矩方法估计第 10 章数据集 B 的参数, 分别假设分布为指数分布, gamma 分布和 Pareto 分布.*

---

① 选择恰当的矩进行估计可以使方程的解落入参数的可行域.

解　样本的前两阶矩分别为

$$\hat{\mu}_1' = \frac{1}{20}(27 + \cdots + 15743) = 1\ 424.4,$$

$$\hat{\mu}_2' = \frac{1}{20}(27^2 + \cdots + 15743^2) = 13\ 238\ 441.9.$$

对于指数分布, 有

$$\theta = 1\ 424.4.$$

显然有 $\hat{\theta} = 1\ 424.4$.

对于 gamma 分布, 两个等式为

$$\mathrm{E}(X) \ = \alpha\theta = 1\ 424.4,$$

$$\mathrm{E}(X^2) = \alpha(\alpha+1)\theta^2 = 13\ 238\ 441.9.$$

第二个等式除以第一个等式的平方得到

$$\frac{\alpha+1}{\alpha} = 6.524\ 89, \quad 1 = 5.524\ 89\alpha.$$

所以 $\hat{\alpha} = 1/5.524\ 89 = 0.181\ 00$ 和 $\hat{\theta} = 1\ 424.4/0.181\ 00 = 7\ 869.61$.

对于 Pareto 分布, 两个方程为

$$\mathrm{E}(X) \ = \frac{\theta}{\alpha-1} = 1\ 424.4,$$

$$\mathrm{E}(X^2) = \frac{2\theta^2}{(\alpha-1)(\alpha-2)} = 13\ 238\ 441.9.$$

用第一个等式的平方除第二个等式得到

$$\frac{2(\alpha-1)}{(\alpha-2)} = 6.524\ 89,$$

得到解 $\hat{\alpha} = 2.442$ 和 $\hat{\theta} = 1\ 424.4 \times 1.442 = 2\ 053.985$. □

这里的讨论既不保证矩估计方程解的存在性, 也不保证解的唯一性.

定义 12.3　*$\boldsymbol{\theta}$ 分位点匹配估计是指满足下列 p 个方程的任意解*

$$\pi_{g_k}(\boldsymbol{\theta}) = \hat{\pi}_{g_k}, \quad k = 1, 2, \cdots, p,$$

其中 $g_1, g_2, \cdots, g_p$ 是 p 个任取的百分数. 由分位点的定义可知, 上述方程可以写成

$$F(\hat{\pi}_{g_k}/\boldsymbol{\theta}) = g_k, \quad k = 1, 2, \cdots, p.$$

这个估计量的出发点是构造一个模型使其 p 个分位点与数据相匹配 (实际数据的分位点按照经验分布计算). 与矩方法一样, 这里并不保证这些方程一定有解, 也不保证解存在情况下的唯一性. 这个定义的一个问题是对于离散型随机变量 (例如经验分布给出的随机变量) 分位点并不总是有明确的定义. 例如, 数据集 B 有 20 个观测值, 在 384 到 457 之间的任何一个数, 都满足有 10 个观测值大于它 10 个观测值小于它, 因此可以作为中位数, 但习惯上我们使用这个区间的中点. 但是对于其他的分位点, 很难有 "正式的" 插值方法[①]. 这里将用到下面的定义.

定义 12.4 分位点的光滑经验估计为

$$\hat{\pi}_g = (1-h)x_{(j)} + hx_{(j+1)},$$
$$j = \lfloor (n+1)g \rfloor \text{ 且 } h = (n+1)g - j.$$

其中 $\lfloor \cdot \rfloor$ 表示最大整数函数, $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$ 为样本的次序统计量.

除非样本中有两个或者两个以上的数据点有相同的值, 任何两个分位点的值都不相同. 这个定义的一个特点是无法获得 $g < 1/(n+1)$ 或者 $g > n/(n+1)$ 情况下的 $\hat{\pi}_g$. 这样看起来是合理的, 因为对于小样本我们不应期望能够得到较大或者较小的分位点. 我们将利用光滑化的经验分布函数进行分位点的经验估计.

例 12.5 采用分位点匹配对数据集 B 进行参数估计, 分别考虑指数分布和 Pareto 分布.

解 对于指数分布, 中位数的经验估计为: $\hat{\pi}_{0.5} = (384+457)/2 = 420.5$, 参数方程为

$$0.5 = F(420.5|\theta) = 1 - \mathrm{e}^{-420.5/\theta},$$
$$\ln 0.5 = \frac{-420.5}{\theta},$$
$$\hat{\theta} = \frac{-420.5}{\ln 0.5} = 606.65.$$

对于 Pareto 分布, 选取 30%和 80%分位点. 光滑化的经验分布估计如下

30%分位点: $j = \lfloor 21(0.3) \rfloor = \lfloor 6.3 \rfloor = 6, h = 6.3 - 6 = 0.3,$

$\hat{\pi}_{0.3} = 0.7(161) + 0.3(243) = 185.6,$

80%分位点: $j = \lfloor 21(0.8) \rfloor = \lfloor 16.8 \rfloor = 16, h = 16.8 - 16 = 0.8,$

$\hat{\pi}_{0.8} = 0.2(1\ 193) + 0.8(1\ 340) = 1\ 310.6.$

参数方程为

① Hyndman & Fan[65] 介绍了 9 种不同的方法. 他们对出现在这里的这个公式进行了一些修正: $j = \lfloor g(n+1/3) + 1/3 \rfloor$ 和 $h = g(n+1/3) + 1/3 - j$.

$$0.3 = F(185.6) = 1 - \left(\frac{\theta}{185.6+\theta}\right)^{\alpha},$$

$$0.8 = F(1\ 310.6) = 1 - \left(\frac{\theta}{1\ 310.6+\theta}\right)^{\alpha},$$

$$\ln 0.7 = -0.356\ 675 = \alpha \ln\left(\frac{\theta}{185.6+\theta}\right),$$

$$\ln 0.2 = -1.609\ 438 = \alpha \ln\left(\frac{\theta}{1\ 310.6+\theta}\right),$$

$$\frac{-1.609\ 438}{-0.356\ 675} = 4.512\ 338 = \frac{\ln(\frac{\theta}{1\ 310.6+\theta})}{\ln(\frac{\theta}{185.6+\theta})}.$$

利用附录 F 中的任何一种方法都可以得到解: $\hat{\theta} = 715.03$. 从而, 由第一个方程, 有

$$0.3 = 1 - \left(\frac{715.03}{185.6 + 715.03}\right)^{\alpha},$$

得到 $\hat{\alpha} = 1.545\ 59$. □

这些估计结果和例 12.2 得到的估计结果有很大的差异, 这也说明这些方法并不特别可靠.

习题

12.1 使用矩方法估计数据集 B 在 250 删失情况下采用指数分布模型的参数.

12.2 使用矩方法估计数据集 B 采用对数正态分布模型的参数.

12.3* 已知某样本的 20%和 80%分位点分别为 5 和 12. 使用分位点匹配方法估计 $S(8)$, 假设总体服从 Weibull 分布.

12.4* 已知某样本的均值为 35 000, 标准差为 75 000, 中位数为 10 000, 90%分位点为 100 000. 使用分位点匹配方法, 估计 Weibull 分布的参数.

12.5* 现有样本: 4, 5, 21, 99 和 421. 用矩方法拟合 Pareto 分布, 并给出拟合分布的 95%分位数.

12.6* 已知如下的索赔记录: 第 1 年有 100 个索赔, 平均索赔额为 10 000; 第 2 年有 200 个索赔, 平均索赔额为 12 500. 通货膨胀使得索赔额每年增长 10%, 采用参数 $\alpha = 3$、θ 未知的 Pareto 分布对索赔额建模. 试用矩方法估计第 3 年的 θ.

12.7* 已知某随机样本的 20%分位点为 18.25, 80%分位点为 35.8. 使用分位点匹配方法估计对数正态分布的参数, 并采用估计结果计算得到一个大于 30 的观测值的概率.

12.8* 某索赔额为两个随机变量 A 和 B 的混合, A 服从均值为 1 的指数分布, B 服从均值为 10 的指数分布. 分布 A 的权重为 p, 分布 B 的权重为 $1-p$. 混合分布的标准差为 2. 试用矩方法估计参数 p.

12.9* 现有 20 个随机样本观测值的排序结果

12	16	20	23	26	28	30	32	33	35
36	38	39	40	41	43	45	47	50	57

试用光滑化经验估计确定样本的 60%分位点.

12.10* 以下为一年内的风灾损失记录 (单位: 百万美元):

1	1	1	1	1	2	2	3	3	4
6	6	8	10	13	14	15	18	22	25

使用光滑化经验估计确定样本的 75%分位点.

12.11* 现有参数 α 和 β 未知的 gamma 分布的观测值: 1 000, 850, 750, 1 100, 1 250, 900, 试用矩方法进行参数估计.

12.12* 现有来自对数 logistic 分布的索赔随机样本, 其中 80% 的索赔超过 100, 20% 的索赔超过 400. 试用分位点匹配方法估计该分布的参数.

12.13* 已知来自累积分布函数 $F(x) = x^p, 0 < x < 1$ 的样本: $x_1, \cdots, x_n$, 试用矩方法估计参数 p.

12.14* 已知服从 gamma 分布的 10 个索赔

1 500　6 000　3 500　3 800　1 800　5 500　4 800　4 200　3 900　3 000

试用矩方法估计参数 α 和 θ.

12.15* 已知服从对数正态分布的 5 个索赔

500　1 000　1 500　2 500　4 500.

试用矩方法估计参数 μ 和 σ, 并估计损失超过 4 500 的概率.

12.16* 随机变量 X 的概率密度函数为 $f(x) = \beta^{-2}x\exp(-0.5x^2/\beta^2)$, $x, \beta > 0$. 对于这个随机变量 $\mathrm{E}(X) = (\beta/2)\sqrt{2\pi}$ 和 $\mathrm{Var}(X) = 2\beta^2 - \pi\beta^2/2$. 现有如下 5 个观测值

4.9　1.8　3.4　6.9　4.0

试用矩方法估计参数 β.

12.17 随机变量 X 的概率密度函数为 $f(x) = \alpha\lambda^\alpha(\lambda+x)^{-\alpha-1}$, $x, \alpha, \lambda > 0$. 已知 $\lambda = 1\,000$, 5 个观测值如下

43　145　233　396　775

试用矩方法估计参数 α.

12.18 试用矩方法和表 12-1 中的数据估计负二项分布模型的参数.

表 12-1　习题 12 .18 的数据

索赔数	保单数
0	9 048
1	905
2	45
3	2
4+	0

12.19 试用矩方法和表 12-2 中的数据估计负二项分布模型的参数.

表 12-2 习题 12.19 的数据

索赔数	保单数
0	861
1	121
2	13
3	3
4	1
5	0
6	1
7+	0

12.2 最大似然估计

12.2.1 引言

使用矩方法和分位点匹配估计通常来说比较容易操作, 但是这些估计量通常表现的不理想. 主要原因是这些方法只用到了数据的部分性质, 没有能够充分地利用全部数据. 当总体分布的右尾很厚时, 是否能够尽可能充分地利用数据的信息显得尤为重要. 例如, 对于正态分布的参数估计, 样本的均值和方差是充分的[①]. 但是, 在估计 Pareto 分布的参数时, 为了正确的估计参数 α 需要掌握所有的极值观测. 这些方法的另外一个缺点是要求所有的观测来自同一个随机变量, 否则很难确定如何用数据来描述总体的矩或分位点. 例如, 当一半的观测来自 50 免赔额的业务, 而另一半观测来自 100 免赔额的业务时, 我们就无法确认样本均值的含义[②]. 最后, 这些方法允许人们任意选取矩的阶数和分位数.

基于个体数据的估计方法很多. 这些估计方法都是事先设定一个目标函数然后通过确定参数的值来优化这个函数. 例如, 若要问题复杂化, 可以首先定义优化目标为: 参数分布函数和 Nelson-Åalen 估计分布函数之差绝对值的最大值, 然后估计参数使得这个优化目标最小化. 在众多的选择中, 这里仅仅考虑最大似然估计方法. 下面将首先介绍这个估计方法的一般形式, 然后是一些有意义的例子.

为了定义最大似然估计量, 设数据集由 n 个事件 $A_1, \cdots, A_n$ 组成, 其中 A_j 为第 j 个观测, 无论观测的形式如何. 例如, A_j 可能是一个点或者一个区间, 后者代表分组数据或截断数据. 例如, 在 u 处删失的观测可能代表发生在 u 到正无穷

① 这时按照统计学的充分性定义和人们的一般常识都是成立的. 如果总体服从正态分布, 则样本均值和方差提供的信息将能够完全代替原始数据.

② 弥补这个缺点的一种方法是首先给定一个数据依赖的模型, 如 Kaplan-Meier 估计, 然后使用该模型的分位点和矩.

区间的事件. 进一步假设事件 A_j 是通过观测随机变量 X_j 得到的, 而且无需要求 $X_1,\cdots,X_n$ 具有相同的分布, 但是要求其分布依赖同一个参数向量 $\boldsymbol{\theta}$, 并且这些随机变量是相互独立的.

定义 12.6 *似然函数的定义为*

$$L(\boldsymbol{\theta})=\prod_{j=1}^{n}\Pr(X_j\in A_j|\boldsymbol{\theta}).$$

*并且, $\boldsymbol{\theta}$ 的最大似然估计是使似然函数取最大值的向量*①.

上述定义并不能保证在符合模型要求的参数值范围内似然函数一定会取得最大值, 并且当某些参数趋于 0 或者无穷的时候似然函数可能会持续增加. 最大化似然函数时一定要特别小心, 因为除了全局最大值外, 可能还存在局部最大值. 通常情况下, 很难解析地求得似然函数的最大值 (通过求偏导数使其等于零). 经常会采用数值求解方法, 尤其是附录 F 中介绍的方法.

由于假设这些观测是相互独立的, 定义中的乘积表示联合概率 $\Pr(X_1\in A_1,\cdots,X_n\in A_n|\boldsymbol{\theta})$, 也就是说, 似然函数表示的是在给定参数下得到观测数据的概率. 最大似然估计的参数使得实际观测到这些数据的可能性最大. 这个估计量的吸引力之一是它几乎在所有情况下都可行. 也就是说, 如果能够给出所求概率的表达式, 就可以应用这个方法. 如果不能用你所选的模型写出或者估算出这样的表达式, 就没有必要考虑这个模型, 因为这个模型并不适用你的问题.

例 12.7 *假设数据集 B 中的数据在 250 处删失, 求指数分布的参数 θ 的最大似然估计.*

解 前 7 个数据点没有删失, 因此集合 A_j 只包含观测值 x_j. 连续模型在某个点的似然函数可这样定义: $\Pr(X_j=x_j)=f(x_j)$. 即似然函数为密度函数. 因此该乘积的前 7 项为

$$f(27)f(82)\cdots f(243)=\theta^{-1}\mathrm{e}^{-27/\theta}\theta^{-1}\mathrm{e}^{-82/\theta}\cdots\theta^{-1}\mathrm{e}^{-243/\theta}=\theta^{-7}\mathrm{e}^{-909/\theta}.$$

对于最后的 13 项, 集合 A_j 是从 250 到正无穷的区间, 因此 $\Pr(X_j\in A_j)=\Pr(X_j>250)=\mathrm{e}^{-250/\theta}$. 有 13 项的形式都是如此, 从而得到这组样本的似然函数为

$$L(\theta)=\theta^{-7}\mathrm{e}^{-909/\theta}(\mathrm{e}^{-250/\theta})^{13}=\theta^{-7}\mathrm{e}^{-4\ 159/\theta}.$$

考虑似然函数对数的最大化更容易一些. 因为常常进行这种处理, 所以记**对数似然**

① 有些作者将似然函数表示为 $L(\theta|\boldsymbol{x})$, 其中的向量 $\boldsymbol{x}$ 表示观测数据. 因为观测数据本身可能具有多种形式, 这个表达式突出了似然函数对观测的依赖性.

函数为 $l(\boldsymbol{\theta})=\ln L(\boldsymbol{\theta})$. 则

$$
\begin{aligned}
l(\theta) &= -7\ln\theta - 4\ 159\theta^{-1}, \\
l'(\theta) &= -7\theta^{-1} + 4\ 159\theta^{-2} = 0, \\
\hat{\theta} &= \frac{4\ 159}{7} = 594.14.
\end{aligned}
$$

在这种情况下, 可以采用微积分的技巧, 令函数的一阶导数等于零. 同时二阶导数在解点的值为负数, 从而验证了函数在这个解点处取得最大值. □

12.2.2 完全的个体数据

当没有截断也没有删失, 并且每一个观测都被记录的时候, 很容易写出对数似然函数.

$$
L(\boldsymbol{\theta}) = \prod_{j=1}^{n} f_{X_j}(x_j/\boldsymbol{\theta}), \quad l(\boldsymbol{\theta}) = \sum_{j=1}^{n} \ln f_{X_j}(x_j|\boldsymbol{\theta}).
$$

这个表达式表明并不要求观测都来自同一个分布.

例 12.8 **使用数据集 B 求如下分布的最大似然估计: 指数分布, 已知 $\alpha=2$ 的 gamma 分布, 两个参数均未知的 gamma 分布.**

解 指数分布的一般解为

$$
\begin{aligned}
l(\theta) &= \sum_{j=1}^{n}(-\ln\theta - x_j\theta^{-1}) = -n\ln\theta - n\bar{x}\theta^{-1}, \\
l'(\theta) &= -n\theta^{-1} + n\bar{x}\theta^{-2} = 0, \\
n\theta &= n\bar{x}, \\
\hat{\theta} &= \bar{x}.
\end{aligned}
$$

对于数据集 B, $\hat{\theta}=\bar{x}=1\ 424.4$. 对数似然函数的值为 -165.23. 在这种情形下, 矩估计和最大似然估计是相等的.

对于 $\alpha=2$ 的 gamma 分布, 有

$$
\begin{aligned}
f(x|\theta) &= \frac{x^{2-1}\mathrm{e}^{-x/\theta}}{\Gamma(2)\theta^2} = x\theta^{-2}\mathrm{e}^{-x/\theta}, \\
\ln f(x|\theta) &= \ln x - 2\ln\theta - x\theta^{-1}, \\
l(\theta) &= \sum_{j=1}^{n}\ln x_j - 2n\ln\theta - n\bar{x}\theta^{-1}, \\
l'(\theta) &= -2n\theta^{-1} + n\bar{x}\theta^{-2} = 0, \\
\hat{\theta} &= \frac{1}{2}\bar{x}.
\end{aligned}
$$

对数据集 B 有：$\hat{\theta} = 1\ 424.4/2 = 712.2$, 对数似然函数值为 -179.98. 同样地, 这个估计与矩估计结果相同.

对于参数未知的 gamma 分布, 方程没有上面那么简单,

$$\begin{aligned} f(x|\alpha,\theta) &= \frac{x^{\alpha-1}\mathrm{e}^{-x/\theta}}{\Gamma(\alpha)\theta^{\alpha}}, \\ \ln f(x|\alpha,\theta) &= (\alpha-1)\ln x - x\theta^{-1} - \ln\Gamma(\alpha) - \alpha\ln\theta. \end{aligned}$$

对 α 的偏导数需要用到 gamma 函数的导数, 导致方程不能得到解析解. 使用数值方法, 得到的估计值为 $\hat{\alpha} = 0.556\ 16$ 和 $\hat{\theta} = 2\ 561.1$, 对数似然函数值为 -162.29. 这个结果与矩估计不同. □

12.2.3 完全的分组数据

当数据是完全的并且是分组形式时, 可以对观测值进行如下归纳. 首先选取一组数 $c_0 < c_1 < \cdots < c_k$, 这里 c_0 是最小的可能观测值 (通常为 0), c_k 是最大的可能观测值 (通常为正无穷). 设 n_j 为落入区间 $(c_{j-1}, c_j]$ 的观测数目, 这种数据的似然函数为

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{k}[F(c_j|\boldsymbol{\theta}) - F(c_{j-1}|\boldsymbol{\theta})]^{n_j},$$

其对数为

$$l(\boldsymbol{\theta}) = \sum_{j=1}^{k} n_j \ln[F(c_j|\boldsymbol{\theta}) - F(c_{j-1}|\boldsymbol{\theta})].$$

例 12.9 *确定数据集 C 基于指数分布的最大似然估计.*

解 对数似然函数为

$$\begin{aligned} l(\theta) &= 99\ln[F(7\ 500) - F(0)] + 42\ln[F(17\ 500) - F(7\ 500)] + \cdots \\ &\quad + 3\ln[1 - F(300\ 000)] \\ &= 99\ln(1 - \mathrm{e}^{-7\ 500/\theta}) + 42\ln(\mathrm{e}^{-7\ 500/\theta} - \mathrm{e}^{-17\ 500/\theta}) + \cdots \\ &\quad + 3\ln \mathrm{e}^{-300\ 000/\theta}. \end{aligned}$$

执行数值计算的程序得到：$\hat{\theta} = 29\ 721$ 和似然函数的值为 -406.03.

12.2.4 截断或删失数据

对于删失数据, 情况并没有变得更加复杂. 如例 12.7 所示, 右删失数据产生了删失点到正无穷的区间. 在例 12.7 中, 删失点以下的数据为个体数据, 因此似然函数中既有密度函数也有分布函数.

截断数据则带来了一些挑战. 一般有两种处理方法. 一种是通过将每个观测值减去截断点对数据平移. 另一种方法认为截断点下方的数据对模型拟合没有带来任何信息.

例 12.10 假设数据集 B 中的值在 200 处从下方截断. 使用两种方法, 估计已知 $\theta=800$ 的 Pareto 分布的参数 α. 然后使用这个模型去估计在免赔额分别为 0, 200, 400 时的平均支付额.

解 使用数据平移方法, 这些值变为: 43, 94, 140, 184, 257, 480, 655, 677, 774, 993, 1 140, 1 684, 2 358, 15 543. 似然函数为

$$\begin{aligned}
L(\alpha) &= \prod_{j=1}^{14} \frac{\alpha(800^{\alpha})}{(800+x_j)^{\alpha+1}}, \\
l(\alpha) &= \sum_{j=1}^{14} [\ln\alpha + \alpha\ln 800 - (\alpha+1)\ln(x_j+800)] \\
&= 14\ln\alpha + 93.584\,6\alpha - 103.969(\alpha+1) \\
&= 14\ln\alpha - 103.969 - 10.384\alpha, \\
l'(\alpha) &= 14\alpha^{-1} - 10.384, \\
\hat{\alpha} &= \frac{14}{10.384} = 1.384\,2.
\end{aligned}$$

因为数据已经被移动了, 所以无法估计无免赔时的成本. 当免赔额为 200 时, 期望支付额为 Pareto 分布的期望值: 800/0.348 2=2 298. 当免赔额上升到 400 时, 等价于对估计的模型设置免赔额 200. 由定理 5.13, 每次支付的期望成本为

$$\frac{\mathrm{E}(X)-\mathrm{E}(X\wedge 200)}{1-F(200)} = \frac{\frac{800}{0.348\,2}\left(\frac{800}{200+800}\right)^{0.348\,2}}{\left(\frac{800}{200+800}\right)^{1.348\,2}} = \frac{1\,000}{0.384\,2} = 2\,872.$$

如果不对数据进行平移, 在建立似然函数时要解决的关键问题是对观测小于 200 的那些被忽略的观测值如何设置概率. 这变成了一种条件概率, 因此似然函数为 (这里的 x_j 为原始数据值)

$$\begin{aligned}
L(\alpha) &= \prod_{j=1}^{14} \frac{f(x_j|\alpha)}{1-F(200|\alpha)} = \prod_{j=1}^{14} \left[\frac{\alpha(800^{\alpha})}{(800+x_j)^{\alpha+1}} \Big/ \left(\frac{800}{800+200}\right)^{\alpha}\right] \\
&= \prod_{j=1}^{14} \frac{\alpha(1,000^{\alpha})}{(800+x_j)^{\alpha+1}}, \\
l(\alpha) &= 14\ln\alpha + 14\alpha\ln 1\,000 - (\alpha+1)\sum_{j=1}^{14}\ln(800+x_j) \\
&= 14\ln\alpha + 96.709\alpha - (\alpha+1)105.810, \\
l'(\alpha) &= 14\alpha^{-1} - 9.101, \\
\hat{\alpha} &= 1.538\,3.
\end{aligned}$$

这是一个无免赔额的模型, 因此没有免赔额的期望成本为 800/0.538 3=1 486. 令免赔额为 200 和 400 得到如下的结果:

$$\frac{\mathrm{E}(X)-\mathrm{E}(X\wedge 200)}{1-F(200)}=\frac{1\ 000}{0.538\ 3}=1\ 858,$$

$$\frac{\mathrm{E}(X)-\mathrm{E}(X\wedge 400)}{1-F(400)}=\frac{1\ 200}{0.538\ 3}=2\ 229.$$ □

现在我们清楚地知道, 大多数的观测都对似然函数有所贡献, 下面用两步概括上述过程.

(1) 对于分子, 如果已知 x 的确切值, 取 $f(x)$; 如果仅仅知道观测值在 y 和 z 之间, 取 $F(z)-F(y)$.

(2) 对于分母, 设 d 是截断点 (如果不是截断数据, 设为 0), 则分母取为 $1-F(d)$.

例 12.11 对于数据集 D2 估计基于 Pareto 和 gamma 分布的生存时间模型.

解 表 12-3 说明了应用这些数值构造似然函数的过程. 对死亡情形时间已知, 所以已知 x 的确切值. 对退保和时刻 5 满期的情形, 观测值是删失的. 因此只知道死亡在退保时刻 y 至无穷区间内的某个时刻发生. 在表格中, $z=\infty$ 没有标记, 因为所有的区间观测在无穷处结束. 对于 Pareto 分布似然函数方程没有显式解, 而且似然函数随着 α 和 θ 的增大不断增加[①], 所以必须通过数值方法进行最大化求解. 对于 gamma 分布, 最大值在 $\hat{\alpha}=2.617$ 和 $\hat{\theta}=3.311$ 处取到. □

表 12-3 例 12.11 的似然函数

观测样本	x,y	d	L	观测样本	x,y	d	L
1	$y=0.1$	0	$1-F(0.1)$	12	$y=3.9$	0	$1-F(3.9)$
2	$y=0.5$	0	$1-F(0.5)$	13	$y=4.0$	0	$f(4.0)$
3	$y=0.8$	0	$1-F(0.8)$	14	$y=4.0$	0	$1-F(4.0)$
4	$x=0.8$	0	$f(0.8)$	15	$y=4.1$	0	$1-F(4.1)$
5	$y=1.8$	0	$1-F(1.8)$	16	$x=4.8$	0	$f(4.8)$
6	$y=1.8$	0	$1-F(1.8)$	17	$y=4.8$	0	$1-F(4.8)$
7	$y=2.1$	0	$1-F(2.1)$	18	$y=4.8$	0	$1-F(4.8)$
8	$y=2.5$	0	$1-F(2.5)$	19~30	$y=5.0$	0	$1-F(5.0)$
9	$y=2.8$	0	$1-F(2.8)$	31	$y=5.0$	0.3	$\frac{1-F(5.0)}{1-F(0.3)}$
10	$x=2.9$	0	$f(2.9)$	32	$y=5.0$	0.7	$\frac{1-F(5.0)}{1-F(0.7)}$
11	$x=2.9$	0	$f(2.9)$	33	$x=4.1$	1.0	$\frac{f(4.1)}{1-F(1.0)}$

① 对于 Pareto 分布, 若参数 α 与 θ 的比例保持不变为常数, 并且两者同时趋于无穷时, 极限分布为指数分布. 因此, 对于这个例子, 指数分布是相对于 Pareto 模型的一个更好的模型 (用似然函数来度量).

(续)

观测样本	x,y	d	L	观测样本	x,y	d	L
34	$x=3.1$	1.8	$\frac{f(3.1)}{1-F(1.8)}$	38	$x=4.0$	3.2	$\frac{f(4.0)}{1-F(3.2)}$
35	$y=3.9$	2.1	$\frac{1-F(3.9)}{1-F(2.1)}$	39	$y=5.0$	3.4	$\frac{1-F(5.0)}{1-F(3.4)}$
36	$y=5.0$	2.9	$\frac{1-F(5.0)}{1-F(2.9)}$	40	$y=5.0$	3.9	$\frac{1-F(5.0)}{1-F(3.9)}$
37	$y=4.8$	2.9	$\frac{1-F(4.8)}{1-F(2.9)}$				

离散数据不会带来额外的问题.

例 12.12 对于数据集 A, 假设 7 个事故次数超过 5 次的司机的实际事故数为 5. 分别假设模型为 Poisson 分布和参数 $m=8$ 的二项分布, 确定最大似然估计.

解 一般来说, 对于具有完全数据的离散分布, 似然函数为

$$L(\boldsymbol{\theta})=\prod_{j=1}^{\infty}[p(x_j|\boldsymbol{\theta})]^{n_j},$$

其中 x_j 为一个观测值, $p(x_j|\boldsymbol{\theta})$ 是观测到 x_j 的概率, n_x 是样本中观测到 x 的次数. 对于 Poisson 分布, 有

$$\begin{aligned}
L(\lambda)&=\prod_{x=0}^{\infty}\left(\frac{\mathrm{e}^{-\lambda}\lambda^x}{x!}\right)^{n_x}=\prod_{x=0}^{\infty}\frac{\mathrm{e}^{-n_x\lambda}\lambda^{xn_x}}{(x!)^{n_x}},\\
l(\lambda)&=\sum_{x=0}^{\infty}(-n_x\lambda+xn_x\ln\lambda-n_x\ln x!)=-n\lambda+n\bar{x}\ln\lambda-\sum_{x=0}^{\infty}n_x\ln x!,\\
l'(\lambda)&=-n+\frac{n\bar{x}}{\lambda}=0,\\
\hat{\lambda}&=\bar{x}.
\end{aligned}$$

对于二项分布, 有

$$\begin{aligned}
L(q)&=\prod_{x=0}^{m}\left[\binom{m}{x}q^x(1-q)^{m-x}\right]^{n_x}=\prod_{x=0}^{m}\frac{m!^{n_x}q^{xn_x}(1-q)^{(m-x)n_x}}{(x!)^{n_x}[(m-x)!]^{n_x}},\\
l(q)&=\sum_{x=0}^{m}[n_x\ln m!+xn_x\ln q+(m-x)n_x\ln(1-q)]\\
&\quad-\sum_{x=0}^{m}[n_x\ln x!+n_x\ln(m-x)!],\\
l'(q)&=\sum_{x=0}^{m}\frac{xn_x}{q}-\frac{(m-x)n_x}{1-q}=\frac{n\bar{x}}{q}-\frac{mn-n\bar{x}}{1-q}=0,\\
\hat{q}&=\frac{\bar{x}}{m}.
\end{aligned}$$

在这里, $\bar{x}=[81\,714(0)+11\,306(1)+1\,618(2)+250(3)+40(4)+7(5)]/94\,935=0.163\,13$. 因此对于 Poisson 过程 $\hat{\lambda}=0.163\,13$, 对于二项分布 $\hat{q}=0.163\,13/8=0.020\,39$. □

习题 12.25 将考虑 5 次或 5 次以上事故数的确切值未知的情况下如何估计 Poisson 模型的参数.

习题

12.20 依次使用逆指数分布, $\alpha=2$ 的逆 gamma 分布和逆 gamma 分布重新计算例 12.8 的结果. 并将你的估计结果与矩估计结果进行比较.

12.21 确定数据集 C 的最大似然估计, 分别假设分布为 gamma 分布, 逆指数分布和逆 gamma 分布.

12.22 确定数据集 B 的最大似然估计, 依次使用逆指数分布, gamma 分布和逆 gamma 分布. 假设数据在 250 删失, 将你得到的答案和例 12.8 及习题 12.20 的结果进行比较.

12.23 使用参数均未知的 Pareto 分布重新计算例 12.10.

12.24 重做例 12.11, 并估计退保时间的分布.

12.25 重做例 12.12, 但假设 7 个事故数超过 5 次的司机的实际事故数未知. 注意这是删失数据情形.

12.26* 通过以下观测寿命估计 q_{35}. 10 个人在 35.4 岁时开始被观测, 其中 6 人在 36 岁之前死亡, 另外 4 人 36 岁仍然活着. 另有 20 人在 35 岁开始观测, 其中 8 人在 36 岁之前死亡, 其他 12 人在 36 岁仍然活着. 若已知 35 岁之后生存时间的密度函数为 $f(t)=w, 0\leqslant t\leqslant 1$; 当 $t>1$ 时 $f(t)$ 无定义. 给出 q_{35} 的最大似然估计.

12.27* 已知模型的风险率函数为: $h(t)=\lambda_1, 0\leqslant t<2$ 和 $h(t)=\lambda_2, t\geqslant 2$. 从新生群体中观测到 5 个数据, 见表 12-4, 确定 λ_1 和 λ_2 的最大似然估计.

表 12-4 习题 12.27 的数据

最后观测的年龄	原因
1.7	死亡
1.5	删失
2.6	删失
3.3	死亡
3.5	删失

12.28* 为了估计 q_x, 假设生存时间 x 具有常值密度函数. 在死亡率调查中有 10 个个体在 x 岁开始观测. 其中, $x+0.5$ 岁时, 有 1 人死亡, 1 人离开观测. 给出 q_x 的最大似然估计.

12.29* 已知 10 个个体的生存函数为

$$S(t)=\left(1-\frac{t}{k}\right)^{1/2}, \quad 0\leqslant t\leqslant k,$$

t 是从出生开始的生存时间. 10 个个体从出生开始观测, 在时刻 10 有 2 个个体死亡,

另外 8 个个体退出观测. 确定 k 的最大似然估计.

12.30* 现有 500 个损失的观测, 其中的 5 个损失为: 1 100, 3 200, 3 300, 3 500, 3 900. 对剩下的 495 个损失只知道它们都超过了 4 000. 基于指数分布模型, 确定均值的最大似然估计.

12.31* 现有 100 个个体在 35 岁观测. 其中 15 人在 35.6 岁时离开调查, 10 人在 35 到 35.6 岁之间死亡, 3 人在 35.6 到 36 岁之间死亡, 剩下的 72 人在 36 岁仍然活着. 给出 q_{35} 的有限乘积估计和最大似然估计. 对后一种估计, 假设生存时间在 35 岁至 36 岁之间服从均匀分布.

12.32* 已知生存函数为 $S(t)=1-t/w, 0\leqslant t\leqslant w$. 现通过研究 5 个索赔以确定从报告到结案的时间分布. 5 年后, 其中的 4 个索赔已结案, 时间分别为 1, 3, 4, 4. 精算师 X 采用最大似然法估计 w. 精算师 Y 决定等到所有的索赔结案. 6 年后第 5 个索赔也结案了, 此时, 精算师 Y 也得到了 w 的最大似然估计. 试给出这 2 个估计.

12.33* 有 4 个机动车引擎在开始使用 3 年后首次观测. 随后继续观测了 r 年, 有 3 个引擎在观测到期前报废, 报废时间为 4, 5, 7. 第 4 个引擎直到 $3+r$ 年仍然可运行. 已知生存函数服从 0 到 w 的均匀分布, w 的最大似然估计为 13.67, 试确定 r.

12.34* 现有 10 个索赔观测. 其中 7 个为 (单位: 千)3, 7, 8, 12, 12, 13, 14. 剩下的 3 个索赔在 15 删失, 已知模型的风险率函数为

$$h(t)=\begin{cases}\lambda_1, & 0<t<5,\\ \lambda_2, & 5\leqslant t<10,\\ \lambda_3, & t\geqslant 10.\end{cases}$$

试给出 3 个参数的最大似然估计.

12.35* 现有 5 个观测值 521, 658, 702, 819, 1 217. 假设模型为单参数 Pareto 分布, 分布函数如下

$$F(x)=1-\left(\frac{500}{x}\right)^{\alpha},\quad x>500, \alpha>0.$$

试给出 α 的最大似然估计.

12.36* 现有如下 5 个索赔额观测: 11.0, 15.2, 18.0, 21.0, 25.8. 试给出下面模型中 μ 的最大似然估计:

$$f(x)=\frac{1}{\sqrt{2\pi x}}\exp\left[-\frac{1}{2x}(x-\mu)^2\right],\quad x,\mu>0.$$

12.37* 现有样本量为 5 的随机样本, 来自 $\tau=2$ 的 Weibull 分布. 其中的 2 个观测超过了 50, 剩下 3 个观测分别为 20, 30 和 45. 试给出参数 θ 的最大似然估计.

12.38* Phil 和 Sylvia 是灯泡行业的竞争对手. Sylvia 的广告宣称她的灯泡的使用时间是 Phil 的 2 倍. 你可以测试 20 个 Phil 的灯泡和 10 个 Sylvia 的灯泡, 假设所有灯泡的寿命服从指数分布, 时间单位为小时. 使用最大似然估计, 得到 Phil 和 Sylvia 的参数估计值分别为 $\hat{\theta}_P=1\ 000$ 和 $\hat{\theta}_S=1\ 500$. 使用全部的 30 个观测, 试给出 $\hat{\theta}^*$, 它代表在 Sylvia 宣称 $\theta_S=2\theta_P$ 的前提下 θ_p 的最大似然估计.

12.39* 现有 100 个损失样本, 其中 62 个小于 1 000, 38 个大于 1 000. 考虑均值为 θ 的指数分布, 使用给出的信息, 试给出 θ 的最大似然估计. 现在知道 62 个小于 1 000 的损失

的总和为 28 140, 大于 1 000 的 38 个损失的总和未知. 使用附加的信息, 求 θ 的最大似然估计.

12.40* 下面的值是由 10 个损失的随机样本计算得到:

$$\sum_{j=1}^{10} x_j^{-2} = 0.000\ 336\ 74, \qquad \sum_{j=1}^{10} x_j^{-1} = 0.023\ 999,$$
$$\sum_{j=1}^{10} x_j^{-0.5} = 0.344\ 45, \qquad \sum_{j=1}^{10} x_j^{0.5} = 488.97,$$
$$\sum_{j=1}^{10} x_j = 31\ 939, \qquad \sum_{j=1}^{10} x_j^2 = 211\ 498\ 983.$$

损失服从 Weibull 分布, 参数 $\tau = 0.5$[因此 $F(x) = 1 - \mathrm{e}^{-(x/\theta)^{0.5}}$], 试给出 θ 的最大似然估计.

12.41* 已知对于 1997 年报告的索赔, 在 1997 年 (当年) 结案的索赔数未知, 在 1998 年 (第 1 年) 结案的索赔数为 3, 在 1999 年 (第 2 年) 的数字为 1, 在 1999 年之后结案的索赔数未知. 对于 1998 年报告的索赔, 有 5 个在当年结案, 2 个在第 1 年结案, 1 年之后的结案数目未知. 对于 1999 年报告的索赔中, 4 个在当年结案, 之后的结案数字未知. 设 N 表示选定索赔的结案年数, 假设服从概率函数: $\Pr(N = n) = p_n = (1-p)p^n$, $n = 0, 1, 2, \cdots$. 试给出 p 的最大似然估计.

12.42* 现有 n 个来自于概率密度函数 $f(x) = 2\theta x \exp(-\theta x^2), x > 0$ 的独立观测 $x_1, \cdots, x_n$. 试给出 θ 的最大似然估计.

12.43* 设 $x_1, \cdots, x_n$ 为来自分布函数为 $F(x) = x^p, 0 < x < 1$ 的随机样本, 试给出 p 的最大似然估计.

12.44 现有服从 gamma 分布的 10 个索赔组成的随机样本为

1 500　6 000　3 500　3 800　1 8000　5 500　4 800　4 200　3 900　3 000

(a)* 假设已知 $\alpha = 12$, 试给出 θ 的最大似然估计.

(b) 试给出 α 和 θ 的最大似然估计.

12.45 现有服从对数正态分布的 5 个索赔随机样本

500　1 000　1 500　2 500　4 500

试给出 μ 和 σ 的最大似然估计. 估计损失超过 4 500 的概率.

12.46* 设 $x_1, \cdots, x_n$ 为来自概率密度函数为 $f(x) = \theta^{-1}\mathrm{e}^{-x/\theta}, x > 0$ 的随机样本. 试给出 θ 的最大似然估计.

12.47* 随机变量 X 的概率密度函数为 $f(x) = \beta^{-2} x \exp(-0.5x^2/\beta^2), x, \beta > 0$. 则有: $\mathrm{E}(X) = (\beta/2)\sqrt{2\pi}$ 和 $\mathrm{Var}(X) = 2\beta^2 - \pi\beta^2/2$, 现有如下 5 个观测

4.9　1.8　3.4　6.9　4.0

试给出 β 的最大似然估计.

12.48* 设 $x_1, \cdots, x_n$ 为来自分布函数 $F(x) = 1 - x^{-\alpha}, x > 1, \alpha > 0$ 的随机样本. 试给出 α 的最大似然估计量.

12.49* 随机变量 X 的概率密度函数为 $f(x) = \alpha\lambda^{\alpha}(\lambda+x)^{-\alpha-1}, x, \alpha, \lambda > 0$. 已知 $\lambda = 1\ 000$, 现有如下 5 个观测值

43 145 233 396 775

试给出 α 的最大似然估计量.

12.50 采集到以下的 20 个观测, 欲估计 $\Pr(X > 200)$, 采用 Pareto 单参数模型: $F(x) = 1 - (100/x)^{\alpha}$, $x > 100, \alpha > 0$.

132	149	476	147	135	110	176	107	147	165
135	117	110	111	226	108	102	108	227	102

(a) 试给出 $\Pr(X > 200)$ 的经验估计.

(b) 试给出单参数 Pareto 分布参数 α 的矩估计, 并利用估计结果计算 $\Pr(X > 200)$.

(c) 试给出单参数 Pareto 分布参数 α 的最大似然估计, 并利用参数估计结果计算 $\Pr(X > 200)$.

12.51 表 12-5 表示 250 个损失样本的统计结果, 考虑逆指数分布, 其累积分布函数为 $F(x) = \mathrm{e}^{-\theta/x}$, $x > 0, \theta > 0$. 试给出 θ 的最大似然估计.

表 12-5 习题 12.51 的数据

损失额	观测数目	损失额	观测数目
0~25	5	350~500	17
25~50	37	500~750	13
50~75	28	750~1000	12
75~100	31	1 000~1 500	3
100~125	23	1 500~2 500	5
125~150	9	2 500~5 000	5
150~200	22	5 000~10 000	3
200~250	17	10 000~25 000	3
250~350	15	25 000~	2

12.52 考虑密度函数如下的逆高斯分布

$$f_X(x) = \left(\frac{\theta}{2\pi x^3}\right)^{1/2} \exp\left[-\frac{\theta}{2x}\left(\frac{x-\mu}{\mu}\right)^2\right], \quad x > 0.$$

(a) 证明

$$\sum_{j=1}^{n} \frac{(x_j - \mu)^2}{x_j} = \mu^2 \sum_{j=1}^{n} \left(\frac{1}{x_j} - \frac{1}{\bar{x}}\right) + \frac{n}{\bar{x}}(\bar{x} - \mu)^2,$$

其中 $\bar{x} = (1/n) \sum\limits_{j=1}^{n} x_j$.

(b) 对于一组样本 $(x_1, \cdots, x_n)$, 证明 μ 和 θ 的最大似然估计为

$$\hat{\mu} = \bar{x}, \hat{\theta} = \frac{n}{\sum_{j=1}^{n}\left(\frac{1}{x_j} - \frac{1}{\bar{x}}\right)}.$$

12.53 假设 $X_1, \cdots, X_n$ 独立, 服从正态分布, 均值 $\mathrm{E}(X_j) = \mu$ 和方差 $\mathrm{Var}(X_j) = (\theta m_j)^{-1}$, 其中 $m_j > 0$ 为已知常数. 证明 μ 和 θ 的最大似然估计为

$$\hat{\mu} = \bar{X}, \hat{\theta} = n\left[\sum_{j=1}^{n} m_j (X_j - \bar{X})^2\right]^{-1},$$

其中 $\bar{X} = (1/m)\sum\limits_{j=1}^{n} m_j X_j$ 和 $m = \sum\limits_{j=1}^{n} m_j$.

12.3 方差和区间估计

在一般情况下, 确定像最大似然估计量这样复杂的估计量的方差非常不容易. 但还是有可能对估计方差进行近似. 关键的基础是大多数数理统计教科书上的一个定理. 本节只是讨论了一种特殊情况的版本, 多参数的一般化情形由 [112] 得到, 在此只陈述结论不给出证明. 这里 $L(\theta)$ 代表似然函数, $l(\theta)$ 代表对数似然函数. 所有的结果都假设总体的分布是某个参数分布族的成员.

定理 12.13 *假设概率密度函数 (在离散情况下为概率函数)$f(x;\theta)$ 在 θ 的一个包含真值的区间内满足以下条件 (对于离散情况下面的积分将换作求和).*

(i) $\ln f(x;\theta)$ *对 θ 三次可微.*

(ii) $\int \frac{\partial}{\partial\theta} f(x;\theta)\mathrm{d}x = 0$. *这意味着求导与积分计算可交换, 所以实际上是对常数 1 微分.*[①]

(iii) $\int \frac{\partial^2}{\partial\theta^2} f(x;\theta)\mathrm{d}x = 0$. *二阶导也具有 (ii) 相同的性质.*

(iv) $-\infty < \int f(x;\theta)\frac{\partial^2}{\partial\theta^2}\ln f(x;\theta)\mathrm{d}x < 0$. *保证了积分的存在性, 并且在导数为零点取最大值.*

(v) *存在函数 $H(x)$, 使得 $\int H(x)f(x;\theta)\mathrm{d}x < \infty$, 并且 $\left|\frac{\partial^3}{\partial\theta^3}\ln f(x;\theta)\right| < H(x)$. 保证总体在极值点不致过于稠密.*

则有如下结论成立:

(a) *当 $n \to \infty$ 时, 似然函数方程 $[L'(\theta) = 0]$ 有解的概率趋近于 1.*

(b) *当 $n \to \infty$ 时, 最大似然估计量 $\hat{\theta}_n$ 的分布收敛到正态分布, 均值为 θ、方*

① (ii) 和 (iii) 的积分都是对满足 $f(x;\theta) > 0$ 的 x 计算.

差满足: $I(\theta)\mathrm{Var}(\hat{\theta}_n) \to 1$, 其中

$$I(\theta) = -n\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ln f(X;\theta)\right] = -n\int f(x;\theta)\frac{\partial^2}{\partial\theta^2}\ln f(x;\theta)\mathrm{d}x$$

$$= n\mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\ln f(X;\theta)\right)^2\right] = n\int f(x;\theta)\left(\frac{\partial}{\partial\theta}\ln f(x;\theta)\right)^2\mathrm{d}x.$$

定理的第二个结论表明, 对于任意的 z, 有

$$\lim_{n\to\infty}\Pr\left(\frac{\hat{\theta}_n-\theta}{[I(\theta)]^{-1/2}} < z\right) = \Phi(z).$$

因此 $[I(\theta)]^{-1}$ 为 $\mathrm{Var}(\hat{\theta}_n)$ 的一个有意义的估计. 称 $I(\theta)$ 为**信息量**(有时候更准确的称为 **Fisher信息量**). 由这个结果我们得到最大似然估计量是渐近无偏和相合的. (i)~(v) 给出的条件经常称为 "一般正则性条件", 对于怀疑论者会这样认为 "这些条件通常是成立的, 但是难以验证, 所以只是假设这些条件都是成立的". 这些条件将确保密度函数对参数比较光滑, 并且密度函数本身比较正常①.

上面的叙述假设样本由独立同分布的随机变量观测组成. 更一般的结果用对数似然函数表示

$$I(\theta) = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}l(\theta)\right] = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}l(\theta)\right)^2\right].$$

这里唯一的要求是每个观测的参数值相同.

如果有一个以上的参数, 结论只是改为参数向量的最大似然估计服从渐近的多元正态分布. 这个分布的协方差阵② 通过求一个 (r,s) 处元素由下式给出的矩阵的逆得到,

$$\mathbf{I}(\boldsymbol{\theta})_{rs} = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta_s\partial\theta_r}l(\boldsymbol{\theta})\right] = -n\mathrm{E}\left[\frac{\partial^2}{\partial\theta_s\partial\theta_r}\ln f(X;\boldsymbol{\theta})\right]$$

$$= \mathrm{E}\left[\frac{\partial}{\partial\theta_r}l(\boldsymbol{\theta})\frac{\partial}{\partial\theta_s}l(\boldsymbol{\theta})\right] = n\mathrm{E}\left[\frac{\partial}{\partial\theta_r}\ln f(X;\boldsymbol{\theta})\frac{\partial}{\partial\theta_s}\ln f(X;\boldsymbol{\theta})\right].$$

其中每行的第一个表达式都是正确的, 第二个表达式假设似然函数是 n 个相同密度函数的乘积. 这个矩阵通常称作**信息阵**. 信息阵也构成了 Cramér-Rao 下界, 也就是说在一般条件下, 任何无偏估计量的方差都大于由信息矩阵的逆给出的方差. 因此, 至少在渐近意义上, 没有哪个无偏估计比最大似然估计更准确.

例 12.14　估计对数正态分布的最大似然估计量的协方差阵. 然后计算数据集 B 的相应结果.

① 习题 12.55 给出了不满足这些条件的一个例子.

② 对于多元随机变量, 协方差阵的主对角线为随机变量的方差, 其余位置为协方差.

解 似然函数和对数似然函数为

$$L(\mu,\sigma)=\prod_{j=1}^{n}\frac{1}{x_j\sigma\sqrt{2\pi}}\exp\left[-\frac{(\ln x_j-\mu)^2}{2\sigma^2}\right],$$

$$l(\mu,\sigma)=\sum_{j=1}^{n}\left[-\ln x_j-\ln\sigma-\frac{1}{2}\ln(2\pi)-\frac{1}{2}\left(\frac{\ln x_j-\mu}{\sigma}\right)^2\right].$$

它们的一阶导数为

$$\frac{\partial l}{\partial\mu}=\sum_{j=1}^{n}\frac{\ln x_j-\mu}{\sigma^2}\quad\text{and}\quad\frac{\partial l}{\partial\sigma}=-\frac{n}{\sigma}+\sum_{j=1}^{n}\frac{(\ln x_j-\mu)^2}{\sigma^3}.$$

二阶导数为

$$\frac{\partial^2 l}{\partial\mu^2}=-\frac{n}{\sigma^2},$$

$$\frac{\partial^2 l}{\partial\sigma\partial\mu}=-2\sum_{j=1}^{n}\frac{\ln x_j-\mu}{\sigma^3},$$

$$\frac{\partial^2 l}{\partial\sigma^2}=\frac{n}{\sigma^2}-3\sum_{j=1}^{n}\frac{(\ln x_j-\mu)^2}{\sigma^4}.$$

期望值为 ($\ln X_j$ 服从均值为 μ, 标准差为 σ 的正态分布)

$$\mathrm{E}\left(\frac{\partial^2 l}{\partial\mu^2}\right)=-\frac{n}{\sigma^2},\quad\mathrm{E}\left(\frac{\partial^2 l}{\partial\mu\partial\sigma}\right)=0,\quad\mathrm{E}\left(\frac{\partial^2 l}{\partial\sigma^2}\right)=-\frac{2n}{\sigma^2}.$$

改变符号后求逆可得到协方差阵的估计 (这只是一个估计, 因为定理 12.13 给出的是极限情况下的协方差阵), 有

$$\begin{bmatrix}\frac{\sigma^2}{n} & 0\\ 0 & \frac{\sigma^2}{2n}\end{bmatrix}.$$

对于对数正态分布, 最大似然估计是下面两个方程的解

$$\sum_{j=1}^{n}\frac{\ln x_j-\mu}{\sigma^2}=0\ \text{且}\ -\frac{n}{\sigma}+\sum_{j=1}^{n}\frac{(\ln x_j-\mu)^2}{\sigma^3}=0.$$

由第一个等式得到 $\hat{\mu}=(1/n)\sum\limits_{j=1}^{n}\ln x_j$, 由第二个等式有 $\hat{\sigma}^2=(1/n)\sum\limits_{j=1}^{n}(\ln x_j-\hat{\mu})^2$. 对于数据集 B, 这些值分别为 $\hat{\mu}=6.137\ 9$ 和 $\hat{\sigma}^2=1.930\ 5$ 或者 $\hat{\sigma}=1.389\ 4$. 协方差矩阵需要参数的真值, 能够做到的最好方法是用估计值代替真值

$$\widehat{\mathrm{Var}}(\hat{\mu},\hat{\sigma})=\begin{bmatrix}0.096\ 5 & 0\\ 0 & 0.048\ 3\end{bmatrix}. \tag{12.1}$$

表达式中的多个“帽子”说明这是估计量方差的一个估计. □

主对角线外的零说明这两个参数估计渐近不相关. 在对数正态分布这个特殊情形, 结论对任何容量的样本都是成立的. 因此可以构造一个置信度为 95%的参数真值的置信区间, 在估计值两边的 1.96 个标准差内有

$$\mu\ :\ 6.137\,9 \pm 1.96(0.096\,5)^{1/2} = 6.137\,9 \pm 0.608\,9,$$
$$\sigma\ :\ 1.389\,4 \pm 1.96(0.048\,3)^{1/2} = 1.389\,4 \pm 0.430\,8.$$

为了得到信息阵需要计算导数和期望值, 这通常并不容易. 可以避免这个问题的一个方法是不计算期望值, 直接使用观测数据, 而不是期望值. 为这个结果称**已观测信息量**.

例 12.15 使用已观测信息量估计前例的协方差阵.

解 将观测值代入二阶导数方程得到

$$\frac{\partial^2 l}{\partial \mu^2} = -\frac{n}{\sigma^2} = -\frac{20}{\sigma^2},$$
$$\frac{\partial^2 l}{\partial \sigma \partial \mu} = -2\sum_{j=1}^{n}\frac{\ln x_j - \mu}{\sigma^3} = -2\frac{122.757\,6 - 20\mu}{\sigma^3},$$
$$\frac{\partial^2 l}{\partial \sigma^2} = \frac{n}{\sigma^2} - 3\sum_{j=1}^{n}\frac{(\ln x_j - \mu)^2}{\sigma^4} = \frac{20}{\sigma^2} - 3\frac{792.080\,1 - 245.515\,2\mu + 20\mu^2}{\sigma^4}.$$

代入参数的估计值得到观测信息的负值项

$$\frac{\partial^2 l}{\partial \mu^2} = -10.360\,0,\quad \frac{\partial^2 l}{\partial_\sigma \partial_\mu} = 0,\quad \frac{\partial^2 l}{\partial \sigma^2} = -20.719\,0.$$

改变符号求逆得到了和 (12.1) 式一样的结果, 这是对数正态分布的特点, 对于其他分布并不一定成立. □

有时, 甚至导数都是不存在的. 这时可以采用二阶导数的近似值, 一种合理的近似为

$$\frac{\partial^2 f(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \doteq \frac{1}{h_i h_j}\left[f\left(\boldsymbol{\theta} + \frac{1}{2}h_i\boldsymbol{e}_i + \frac{1}{2}h_j\boldsymbol{e}_j\right) - f\left(\boldsymbol{\theta} + \frac{1}{2}h_i\boldsymbol{e}_i - \frac{1}{2}h_j\boldsymbol{e}_j\right)\right.$$
$$\left. - f\left(\boldsymbol{\theta} - \frac{1}{2}h_i\boldsymbol{e}_i + \frac{1}{2}h_j\boldsymbol{e}_j\right) + f\left(\boldsymbol{\theta} - \frac{1}{2}h_i\boldsymbol{e}_i - \frac{1}{2}h_j\boldsymbol{e}_j\right)\right],$$

其中 $\boldsymbol{e}_i$ 为这种向量: 除了第 i 个位置为 1, 其余位置均为零. $h_i = \theta_i/10^v$, v 为计算中有效数字位数的 1/3.

例 12.16 使用导数的近似计算重做前面的例子.

解 假设计算中采用 15 位有效数字. 则 $h_1 = 6.137\,9/10^5$ 和 $h_2 = 1.3894/10^5$. 与其接近的数为 0.000 06 和 0.000 01. 第一个近似为

$$\begin{aligned}\frac{\partial^2 l}{\partial \mu^2} &\doteq \frac{l(6.137\,96, 1.389\,4) - 2l(6.137\,9, 1.389\,4) + l(6.137\,84, 1.389\,4)}{(0.00006)^2} \\ &= \frac{-157.713\,893\,081\,98 - 2(-157.713\,893\,049\,68) + (-157.713\,893\,054\,68)}{(0.000\,06)^2} \\ &= -10.360\,4.\end{aligned}$$

另外两个近似为

$$\frac{\partial^2 l}{\partial_\sigma \partial_\mu} \doteq 0.000\,3, \quad \frac{\partial^2 l}{\partial \sigma^2} \doteq -20.720\,8.$$

我们看到在这里近似的效果相当不错. □

信息矩阵提供了一种评价参数的最大似然估计质量的方法. 但是, 人们更关心作为参数函数的某些量的估计, 比如, 我们希望以对数正态分布的均值作为总体均值的估计. 也就是说, 以 $\exp(\hat{\mu} + \hat{\sigma}^2/2)$ 作为总体均值的一个估计, 其中采用了最大似然估计量. 计算这个随机变量的均值和方差是很困难的, 因为其中的两个随机变量的分布已经相当的复杂. 下面的定理 (见 [108]) 对此所有帮助, 常称之为**delta 方法**.

定理 12.17 *令 $\boldsymbol{X}_n = (X_{1n}, \cdots, X_{kn})^{\mathrm{T}}$ 表示一个容量为 n 的 k 维多元随机变量样本. 假设 $\boldsymbol{X}$ 为渐近正态分布, 均值为 $\boldsymbol{\theta}$, 协方差阵为 $\sum/n$, 其中 $\boldsymbol{\theta}$ 和 $\sum$ 都不依赖于 n. g 是一个完全可微的 k 元函数, $G_n = g(X_{1n}, \cdots, X_{kn})$. 则 G_n 也是均值为 $g(\boldsymbol{\theta})$、方差为 $(\partial \boldsymbol{g})^{\mathrm{T}} \sum (\partial \boldsymbol{g})/n$ 的渐近正态, 其中 $\partial \boldsymbol{g}$ 为一阶导数向量 $\partial \boldsymbol{g} = (\partial g/\partial \theta_1, \cdots, \partial g/\partial \theta_k)^T$, 而且取值于原随机变量参数的真值 $\boldsymbol{\theta}$.*

定理的陈述比较难解释. X 是估计量, g 为待估参数的函数. 对于单参数模型, 这个定理的陈述如下: 令 $\hat{\theta}$ 为 θ 的估计量, 服从均值为 θ、方差为 σ^2/n 的渐近正态分布. 则 $g(\hat{\theta})$ 也服从渐近正态分布, 均值为 $g(\theta)$, 渐近方差为 $[g'(\theta)](\sigma^2/n)[g'(\theta)] = g'(\theta)^2\sigma^2/n$.

例 12.18 *使用 delta 方法给出指数分布超过 200 的概率值的最大似然估计的方差, 然后计算数据集 B 的结果.*

解 由例 12.8 知, 指数分布参数的最大似然估计是样本均值. 待估计的量为 $p = \Pr(X > 200) = \exp(-200/\theta)$. 其最大似然估计为 $\widehat{p} = \exp(-200/\widehat{\theta}) = \exp(-200/\overline{x})$. 计算这个估计量的均值和方差并不容易, 但是已知 $\mathrm{Var}(\bar{X}) = \mathrm{Var}(X)/n = \theta^2/n$. 进一步有,

$$g(\theta) = \mathrm{e}^{-200/\theta}, \quad g'(\theta) = 200\theta^{-2}\mathrm{e}^{-200/\theta},$$

因此由 delta 方法有

$$\mathrm{Var}(\hat{p}) \doteq \frac{(200\theta^{-2}\mathrm{e}^{-200/\theta})^2\theta^2}{n} = \frac{40,000\theta^{-2}\mathrm{e}^{-400/\theta}}{n}.$$

对于数据集 B, 有

$$\begin{aligned}\bar{x} &= 1\ 424.4,\\ \hat{p} &= \exp\left(-\frac{200}{1\ 424.4}\right) = 0.869\ 00,\\ \widehat{\mathrm{Var}}(\hat{p}) &= \frac{40\ 000(1\ 424.4)^{-2}\exp(-400/1\ 424.4)}{20} = 0.000\ 744\ 4.\end{aligned}$$

p 的 95% 置信区间为 $0.869 \pm 1.96\sqrt{0.000\ 744\ 4}$ 或者 0.869 ± 0.053. □

例 12.19 利用数据集 B 构造对数正态总体均值的 95% 置信区间. 将这个结果与由传统方法样本均值得到的置信区间作比较.

解 由例 12.14 有 $\hat{\mu} = 6.137\ 9$ 和 $\hat{\sigma} = 1.389\ 4$, 两个估计量的协方差矩阵为

$$\frac{\hat{\Sigma}}{n} = \begin{bmatrix} 0.096\ 5 & 0 \\ 0 & 0.048\ 3 \end{bmatrix},$$

函数 $g(\mu,\sigma) = \exp(\mu + \sigma^2/2)$. 偏导数为

$$\frac{\partial g}{\partial \mu} = \exp\left(\mu + \frac{1}{2}\sigma^2\right), \quad \frac{\partial g}{\partial \sigma} = \sigma\exp\left(\mu + \frac{1}{2}\sigma^2\right),$$

这些量的估计值分别为 1 215.75 和 1 689.16. 由 delta 方法得到下面的估计

$$\widehat{\mathrm{Var}}[g(\hat{\mu},\hat{\sigma})] = [1\ 215.75\ \ 1\ 689.16]\begin{bmatrix} 0.096\ 5 & 5 \\ 0 & 0.048\ 3 \end{bmatrix}\begin{bmatrix} 1\ 215.75 \\ 1\ 689.16 \end{bmatrix} = 280\ 444$$

置信区间为 $1\ 215.75 \pm 1.96\sqrt{280\ 444}$ 或 $1\ 215.75 \pm 1\ 037.96$.

按照传统方法, 总体均值的置信区间为样本均值的邻域 $\bar{x} \pm 1.96s/\sqrt{n}$, 这里 s^2 为样本方差. 对于数据集 B 这个区间为 $1\ 424.4 \pm 1.96(3\ 435.04)/\sqrt{20}$ 或者 $1\ 424.4 \pm 1\ 505.47$. 这是一个更大的区间, 我们并不感到惊讶, 因为我们知道 (对于对数正态分布) 最大似然估计量是渐近的 UMVUE. □

习题

12.54 对数据集 B 给出指数分布和 gamma 分布参数的 95% 置信区间. 似然函数和最大似然估计可由例 12.8 得到.

12.55 X 服从 0 到 θ 的均匀分布. 证明 θ 的最大似然估计量为 $\hat{\theta} = \max(X_1, \cdots, X_n)$. 利用例 9.7 和例 9.10 证明该估计量是渐近无偏的, 并求其方差. 证明: 由定理 12.13 得到的方差估计为负, 且定理的条件 (ii) 不满足.

12.56 对数据集 B 使用 delta 方法确定 gamma 分布均值的 95%置信区间, 可利用习题 12.54 的相关结果.

12.57* 已知对数正态分布的最大似然估计为 $\hat{\mu} = 4.215$ 和 $\hat{\sigma} = 1.093$, 估计量 $(\hat{\mu}, \hat{\sigma})$ 的协方差矩阵为

$$\begin{bmatrix} 0.119\,5 & 0 \\ 0 & 0.059\,7 \end{bmatrix}.$$

对数正态分布的均值为 $\exp(\mu + \sigma^2/2)$. 利用 delta 方法估计这个对数正态分布均值的最大似然估计量的方差.

12.58* 已知某分布有两个参数 α 和 β. 现有容量为 10 的样本有如下的对数似然函数

$$l(\alpha, \beta) = -2.5\alpha^2 - 3\alpha\beta - \beta^2 + 50\alpha + 2\beta + k,$$

其中 k 为常数, 试给出最大似然估计量 $(\hat{\alpha}, \hat{\beta})$ 的协方差矩阵.

12.59 在习题 12.39 中对于同一个模型得到了两个最大似然估计. 第二个估计是在比第一个更多的信息的基础上得到的. 有理由认为第二个估计比第一个估计更加准确. 通过计算两个估计量的方差确认此假设, 请使用已观测似然量进行计算.

12.60 本题为习题 12.43 的继续. 令 $x_1, \cdots, x_n$ 为总体的一个样本, 总体分布函数为 $F(x) = x^p, 0 < x < 1$.

(a) 试给出 p 的最大似然估计量的渐近方差.

(b) 利用你得到的结果给出 p 的 95%置信区间的一般公式.

(c) 试给出 $E(X)$ 的最大似然估计和它的渐近方差以及 95%置信区间的计算公式.

12.61 本题为习题 12.46 的继续. 令 $x_1, \cdots, x_n$ 为总体的一个样本, 总体的概率密度函数为 $f(x) = \theta^{-1}\mathrm{e}^{-x/\theta}, x > 0$.

(a) 试给出 θ 的最大似然估计量的渐近方差.

(b)* 利用得到的结果给出 θ 的 95%置信区间的一般公式.

(c) 试给出 $\mathrm{Var}(X)$ 的最大似然估计和它的渐近方差以及 95%置信区间的计算公式.

12.62* 现有容量为 40 来自概率密度函数 $f(x) = (2\pi\theta)^{-1/2}\mathrm{e}^{-x^2/(2\theta)}, -\infty < x < \infty, \theta > 0$ 的样本, θ 的最大似然估计量 $\hat{\theta} = 2$. 试给出 $\hat{\theta}$ 的 MSE.

12.63 现有来自密度函数为 $f(x) = 2\lambda x\mathrm{e}^{-\lambda x^2} (x, \lambda > 0)$ 的 4 个观测. 其中恰有一个小于 2.

(a)* 试给出 λ 的最大似然估计量.

(b) 试给出 λ 的最大似然估计量的方差.

12.64 根据习题 12.44 的数据, 试给出其最大似然估计量的协方差矩阵, 其中 α 和 θ 均未知. 计算对数似然函数导数的估计量, 然后构造均值的 95%置信区间.

12.65 试给出习题 12.49 中最大似然估计量的方差, 并利用它构造 $\mathrm{E}(X \wedge 500)$ 的 95%置信区间.

12.66 考虑一个容量为 n 来自 Weibull 分布的随机样本. Weibull 生存函数为

$$S(x) = \exp\left\{-\left[\frac{\Gamma(1+\tau^{-1})x}{\mu}\right]^{\tau}\right\}.$$

假设 τ 是已知的, 只有 μ 是待估参数.

(a) 证明 $\mathrm{E}(X)=\mu$.

(b) 证明 μ 的最大似然估计为

$$\hat{\mu}=\Gamma(1+\tau^{-1})\left(\frac{1}{n}\sum_{j=1}^{n}x_j^{\tau}\right)^{1/\tau}.$$

(c) 证明：使用观测信息量得到方差的估计为

$$V\hat{a}r(\hat{\mu})=\frac{\hat{\mu}^2}{n\tau^2},$$

其中 μ 被 $\hat{\mu}$ 替代.

(d) 证明：使用信息量 (仍然用 $\hat{\mu}$ 替代 μ) 得到和 (c) 相同的方差估计.

(e) 证明：$\hat{\mu}$ 服从变形的 gamma 分布，参数 $\alpha=n,\theta=\mu n^{-1/\tau}$ 和 $\tau=\tau$. 利用这个结论得到 $\hat{\mu}$(作为 μ 的函数) 的准确方差. 提示：随机变量 X^{τ} 服从指数分布，所以随机变量 $\sum\limits_{j=1}^{n}X_j^{\tau}$ 服从 gamma 分布，第一个参数为 n，第二个参数为指数分布的均值.

12.4　贝叶斯估计

目前为止对估计方法的讨论都是基于频率学派的假设进行的. 即，总体的分布是确定的但是未知，我们的决定不仅取决于已有的样本，还与其他可能得到的样本的概率有关. 贝叶斯方法假设，只有实际的观测数据是相关的，并且将总体看作变量. 下面的定义描述了参数估计过程，然后由贝叶斯定理给出解.

12.4.1　定义和贝叶斯定理

定义 12.20　*先验分布是在参数的可能取值空间上的一个概率分布，记为 $\pi(\theta)$，代表人们对于真值各种取值可能性的看法.*

如前所述，参数 θ 可以是标量也可以是向量. 关于先验分布的假定通常成为贝叶斯方法被广泛接受的障碍. 在得到第一个数据之前，几乎可以肯定你的经验应该已经提供了一些对参数取值的认识. (如果不是这样，那么分配给你任务的人的智慧可能就值得怀疑了.) 困难的是把这种认识转化为一种概率分布. 可以在 Lindley[83] 中找到一个很好的关于先验分布和贝叶斯分析的讨论，关于贝叶斯和频率学派方法比较的讨论，参见 Efron[32]. Klugman[77] 包括了更详细的对贝叶斯方法的讨论，并给出了许多精算的应用. 贝叶斯方法应用于精算的最新文章有 [25],[101],[119],[133]. 一本不错的关于贝叶斯方法数学推导的教材是 Berger[13]. 最近几年，贝叶斯计算有了很多进展，文献 [22] 就是一本不错的参考资料. Scollnik[118] 已经演示了如何使用计算机程序 WINBUGS 计算精算问题的贝叶斯解.

由于很难发现先验分布 (你将不得不说服其他人接受你的先验分布), 并且有时可能并没有关于先验分布的预先认识, 所以先验分布的定义可以放宽.

定义 12.21 称概率 (或者概率密度函数) 值非负, 但是总和 (或全积分) 为无穷的分布为非正常的先验分布.

对于所谓的**无信息**或者**模糊先验**已经有很多的研究, 这些先验分布只反映了很少的信息, 而对构造模糊先验的最好方法还没有达成共识. 但是, 对于尺度参数族存在一个一致认同的合适的无信息先验 $\pi(\theta)=1/\theta,\theta>0$. 注意这也是一个非正常的先验分布.

贝叶斯分析中的模型和以前没有什么不同.

定义 12.22 模型分布是当参数为特定值时所收集数据的概率分布. 它的概率密度函数表示为 $f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}|\theta)$, 其中 $\boldsymbol{x}$ 的向量形式提醒我们它包含了所有的数据. 同样需要注意的是, 这与似然函数相同, 所以也经常使用似然函数.

如果观测向量 $\boldsymbol{x}=(x_1,\cdots,x_n)^{\mathrm{T}}$ 由独立同分布的随机变量组成, 则

$$f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}|\theta)=f_{\boldsymbol{X}|\Theta}(x_1|\theta)\cdots f_{\boldsymbol{X}|\Theta}(x_n|\theta).$$

我们采用多元统计中的概念给出两个新的定义. 在以下的两种情形中, 若分布为离散的都应该将积分用求和来替换.

定义 12.23 联合分布的概率密度函数为

$$f_{\boldsymbol{X},\Theta}(\boldsymbol{x},\theta)=f_{\boldsymbol{x}|\Theta}(\boldsymbol{x}|\theta)\pi(\theta).$$

定义 12.24 $\boldsymbol{x}$ 的边缘分布的概率密度函数为

$$f_{\boldsymbol{X}}(\boldsymbol{x})=\int f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}|\theta)\pi(\theta)\mathrm{d}\theta.$$

将这个定义与 4.4.5 节混合分布的定义 (4.4) 式作比较可以得到我们感兴趣的以下两个量.

定义 12.25 后验分布是对于给定数据的关于参数的条件概率分布, 记为 $\pi_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$.

定义 12.26 预测分布是数据 $\boldsymbol{x}$ 给定时一个新的观测 y 的条件概率分布. 记为 $f_{Y|\boldsymbol{X}}(y|\boldsymbol{x})$.①

这两个定义是贝叶斯分析的主要内容. 后验分布告诉我们当观测发生变化时参数是如何变化的. 预测分布告诉我们, 根据数据的信息 (还隐含了我们的先验观点) 下一个观测将是什么样子. 贝叶斯定理则告诉我们如何计算后验分布.

① 在这一节和后面贝叶斯方法的讨论中, 我们仍然用 $f(\cdot)$ 表示观测 (比如模型和预测) 的分布, $\pi(\cdot)$ 表示参数的分布 (比如先验和后验分布). 其中的变量会清楚的显示出所使用的特定分布, 为了更加明确, 我们用脚标代表对应的随机变量.

定理 12.27　后验分布的计算如下

$$\pi_{\Theta|\boldsymbol{x}}(\theta|\boldsymbol{x})=\frac{f_{\boldsymbol{x}|\Theta}(\boldsymbol{x}|\theta)\pi(\theta)}{\displaystyle\int f_{\boldsymbol{x}|\Theta}(\boldsymbol{x}|\theta)\pi(\theta)\mathrm{d}\theta},\tag{12.2}$$

预测分布的计算如下

$$f_{Y|\boldsymbol{X}}(y|\boldsymbol{x})=\int f_{Y|\Theta}(y|\theta)\pi_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})\mathrm{d}\theta,\tag{12.3}$$

其中 $f_{Y|\Theta}(y|\theta)$ 是给定参数值后，一个新观测的概率密度函数.

预测分布可以解释为一个混合分布，是以后验分布进行混合. 下面的例子具体说明了上面的定义和结果. 背景 (不包括数据) 源自 Meyers[92].

例 12.28　下面的数据为医疗责任保险单的赔付记录

125　132　141　107　133　319　126　104　145　223

每笔赔付服从单参数 Pareto 分布：$\theta=100$、α 未知. 先验分布服从 gamma 分布：$\alpha=2,\theta=1$. 试给出所有相关的贝叶斯量.

解　先验密度为

$$\pi(\alpha)=\alpha\mathrm{e}^{-\alpha},\quad \alpha>0,$$

模型为 (在每个数据点计算)

$$f_{\boldsymbol{X}|A}(\boldsymbol{x}|\alpha)=\frac{\alpha^{10}(100)^{10\alpha}}{\left(\displaystyle\prod_{j=1}^{10}x_j^{\alpha+1}\right)}=\alpha^{10}\mathrm{e}^{-3.801\ 121\alpha-49.852\ 823}.$$

$\boldsymbol{x}$ 和 A 的联合密度为 (同样在数据点计算)

$$f_{\boldsymbol{X},A}(\boldsymbol{x},\alpha)=\alpha^{11}\mathrm{e}^{-4.801\ 121\alpha-49.852\ 823}.$$

则 α 的后验分布为

$$\pi_{A|\boldsymbol{X}}(\alpha|\boldsymbol{x})=\frac{\alpha^{11}\mathrm{e}^{-4.801\ 121\alpha-49.852\ 823}}{\displaystyle\int_0^\infty\alpha^{11}\mathrm{e}^{-4.801\ 121\alpha-49.852\ 823}\mathrm{d}\alpha}=\frac{\alpha^{11}\mathrm{e}^{-4.801\ 121\alpha}}{(11!)(1/4.801\ 121)^{12}}.\tag{12.4}$$

并不需要计算分母中的积分，因为我们知道最终结果必须为概率分布函数，分母只是一个近似的正规化常数，观察分子得到 gamma 分布：$\alpha=12$ 和 $\theta=1/4.801\ 121$.

则预测分布为

$$f_{Y|\boldsymbol{X}}(y|\boldsymbol{x})=\int_0^\infty\frac{\alpha100^\alpha}{y^{\alpha+1}}\frac{\alpha^{11}\mathrm{e}^{-4.801\ 121\alpha}}{(11!)(1/4.801\ 121)^{12}}\mathrm{d}\alpha$$

$$= \frac{1}{y(11!)(1/4.801\ 121)^{12}} \int_0^\infty \alpha^{12} \mathrm{e}^{-(0.195\ 951+\ln y)\alpha} \mathrm{d}\alpha$$

$$= \frac{1}{y(11!)(1/4.801\ 121)^{12}} \frac{(12!)}{(0.195\ 951 + \ln y)^{13}}$$

$$= \frac{12(4.801\ 121)^{12}}{y(0.195\ 951 + \ln y)^{13}}, \quad y > 100. \tag{12.5}$$

这个密度函数看上去并不熟悉, 习题 12.67 将证明 $\ln Y - \ln 100$ 服从 Pareto 分布.□

12.4.2 推断和预测

从某种意义上讲我们的分析工作已经完成了. 始于一个关于参数或下一个观测的已知分布, 最终得到了一个修正的分布. 但是, 很可能你的老板并不会满意于对他或她的问题只给出了一个分布的回答. 无庸置疑, 一个具体的数值, 可能再加上某个误差范围才是他想要的. 通常贝叶斯方法的解决方案是给出一个损失函数.

定义 12.29 *损失函数$l_j(\hat{\theta}_j, \theta_j)$ 将刻画人们采用估计量 $\hat{\theta}_j$ 作为第 j 个参数的真值 θ_j 的估计时, 所产生的损失量.*

损失函数 $l(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$ 可以为多元的, 这时损失程度同时依赖于多个参数的估计误差.

定义 12.30 *基于某个损失函数的贝叶斯估计是使得期望损失最小化的后验分布参数.*

3 种最常用的损失函数定义如下.

定义 12.31 *对于平方误差损失, 损失函数为 (为了简化省略所有的下标)$l(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$. 对于绝对损失, 损失函数为 $l(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$. 对于零损失, 损失函数为: 当 $\hat{\theta} = \theta$ 时, $l(\hat{\theta}, \theta) = 0$; 其他情况等于 1.*

下面的定理给出了 3 种常用的损失函数的贝叶斯估计.

定理 12.32 *在平方误差损失情形, 贝叶斯估计为后验分布的均值; 在绝对损失情形, 贝叶斯估计为中位数; 在零一损失情形, 贝叶斯估计为众数.*

注意定理并不保证后验分布的均值一定存在或者后验分布的中位数和众数是唯一的. 在没有其他的特别说明时, 贝叶斯估计一般表示后验分布的均值.

例 12.33 (续例 12.28) *试给出 α 的 3 个贝叶斯估计.*

解 后验 gamma 分布的均值为 $\alpha\theta = 12/4.801\ 121 = 2.499\ 416$. 中位数 2.430 342 必须通过数值方法得到, 众数为 $(\alpha - 1)\theta = 11/4.801\ 121 = 2.291\ 131$. 注意这里的 α 是后验 gamma 分布的参数, 而不是单参数 Pareto 分布中要估计的 α. □

为了达到预报的目的, 通常是要预测分布的均值. 它可以被看作是给定 n 个观测和先验分布时对第 $n+1$ 个观测的估计, 即

$$\mathrm{E}(Y|\boldsymbol{x}) = \int y f_{Y|\mathbf{X}}(y|\boldsymbol{x}) \mathrm{d}y$$

$$
\begin{aligned}
&= \int y \int f_{Y|\Theta}(y|\theta)\pi_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})\mathrm{d}\theta\mathrm{d}y \\
&= \int \pi_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x}) \int y f_{Y|\Theta}(y|\theta)\mathrm{d}y\mathrm{d}\theta \\
&= \int \mathrm{E}(Y|\theta)\pi_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})\mathrm{d}\theta. \qquad (12.6)
\end{aligned}
$$

方程 (12.6) 可以看作是以后验分布为权重的加权平均.

例 12.34 (续例 12.28) *在给定前 10 个观测的基础上确定第 11 个观测的期望值.*

解 对于单参数 Pareto 分布, 当 $\alpha > 1$ 时, $\mathrm{E}(Y|\alpha) = 100\alpha/(\alpha-1)$. 由于当 $\alpha \leqslant 1$ 时后验分布仍然有正的概率值, 所以这个预测分布的期望无法定义. □

贝叶斯方法也可以很容易的构造与置信区间类似的概念.

定义 12.35 *称满足如下条件的 $a < b$ 为 θ_j 的 $100(1-\alpha)\%$信度区间, 只要 $\Pr(a \leqslant \Theta_j \leqslant b|\boldsymbol{x}) \geqslant 1-\alpha$ 成立.*

使用信度的概念与第 16 章中精算分析的含义没有关系. 这个不等式如此表达是为了适应 θ_j 的后验分布为离散分布的情形, 那时让这个概率精确等于 $1-\alpha$ 或许是不可能的. 这个定义没有说明解的唯一性, 下面的定理提供了构造唯一区间的方法.

定理 12.36 *若后验随机变量 $\theta_j|\boldsymbol{x}$ 为连续单峰的, 则使 $b-a$ 达到最小的 $100(1-\alpha)\%$ 信度区间是唯一的, 且满足*

$$
\begin{aligned}
\int_a^b \pi_{\Theta_j|\boldsymbol{X}}(\theta_j|\boldsymbol{x})\mathrm{d}\theta_j &= 1-\alpha, \\
\pi_{\Theta|\boldsymbol{X}}(a|\boldsymbol{x}) &= \pi_{\Theta|\boldsymbol{x}}(b|\boldsymbol{x}).
\end{aligned}
$$

这个区间是最高后验密度 (HPD) 信度集的一个特例.

下面的例子将具体说明这个定理.

例 12.37 (续例 12.28) *试给出参数 α 的最窄的 95% 信度区间, 同时给出一个信度区间使得在两个端点外分别有 2.5%的概率.*

解 定理 12.36 的两个方程为

$$
\begin{aligned}
&\Pr(a \leqslant A \leqslant b|\boldsymbol{x}) = \Gamma(12; 4.801\ 121b) - \Gamma(12; 4.801\ 121a) = 0.95, \\
&a^{11}\mathrm{e}^{-4.801\ 121a} = b^{11}\mathrm{e}^{-4.801\ 121b}.
\end{aligned}
$$

使用数值方法可以得到 $a = 1.183\ 2$ 和 $b = 3.938\ 4$. 区间的宽为 2.755 2.

若要求两端点外分别为 2.5%的概率, 可得到如下两个方程

$$
\Gamma(12; 4.801\ 121b) = 0.975, \quad \Gamma(12; 4.801\ 121a) = 0.025.
$$

这个解需要使用不完整 gamma 函数的反函数, 或者使用不完整 gamma 函数的求根技巧. 解为 $a = 1.291\ 5$ 和 $b = 4.099\ 5$. 宽度为 2.808 0, 比第一个区间要宽. 图 12-1 显示了这两个区间的差异, 较细的垂线表示 HPD 区间, 在这两条线左端和右端的面积之和为 0.05. 另一个区间也有 95% 的概率. 为了使两边概率均为 0.025, 需要把两条线同时向右移动. 为了使右端减少的概率等于左端增加的概率, 右侧的上限必须移动的多一些. 因为后验密度在右边那段的取值比左端要低. 这样必然得到一个更大的区间. □

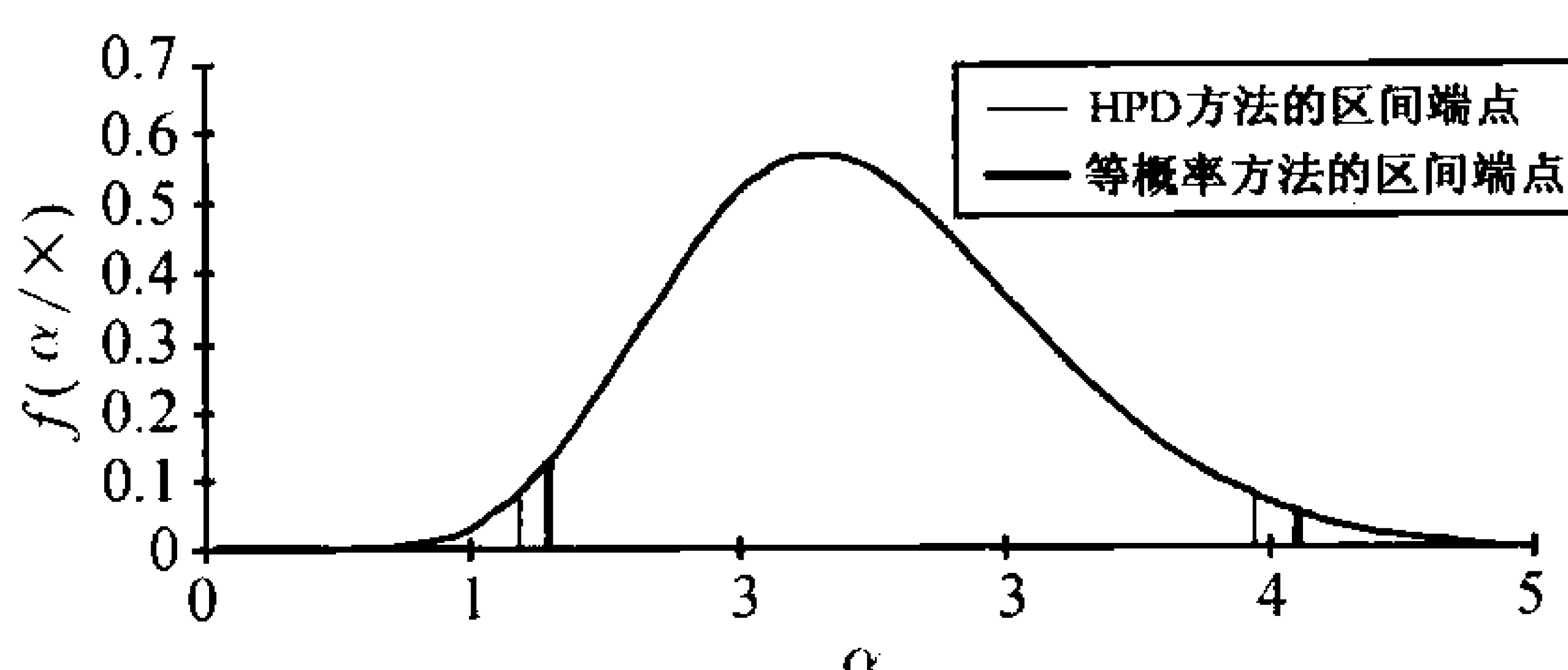


图 12-1 两个贝叶斯信度区间

下面的定义给出了对于任何后验分布都等价的结果.

定义 12.38 对于任何后验概率分布, 称满足以下条件的参数值集合 C 为 $100(1-\alpha)\%$HPD 信度集

$$\Pr(\theta_j \in C) \geqslant 1-\alpha, \tag{12.7}$$

$$C = \{\theta_j : \pi_{\Theta_j|\mathbf{X}}(\theta_j|x) \geqslant c\}, \text{对某些} c,$$

其中 c 是使不等式 (12.7) 成立的最大值.

这个集合可能是很多区间的并集 (当后验分布为多峰的时候将会如此). 这个定义构造了一个宽度最小的满足后验概率的集合. 为了构造这个集合, 首先令 c 等于一个很大的值, 然后逐步减小. 当它减少的时候, 集合 C 变大, 概率也变大, 直到概率达到 $1-\alpha$. 在向量参数 θ 情形, 这个定义可以推广为一个联合的信度集.

有时很难计算后验分布, 但计算后验的矩比较容易. 因此我们可以利用贝叶斯中心极限定理. 下面是来自 Berger[13] (p.224) 的一个解释.

定理 12.39 当 $\pi(\theta)$ 和 $f_{\mathbf{X}|\theta}(\boldsymbol{x}|\theta)$ 对于 θ 都是二次可微的, 并且其他通常需要的假设成立时, 给定 $\boldsymbol{X} = \boldsymbol{x}$ 时 Θ 的后验分布为渐近正态的.

"通常需要的假设" 与定理 12.13 中的描述相似. 与那个定理类似, 这里也可以作进一步的估计. 特别地, 如果把后验分布的众数用后验分布的均值替代并且/或者如果使用后验概率密度的负对数的二阶偏导阵的逆来估计后验协方差矩阵, 渐近正态分布仍然成立.

例 12.40 (续例 12.28) 利用贝叶斯中心极限定理构造 α 的 95% 信度区间.

解 后验分布的均值为 2.499 416, 方差为 $\alpha\theta^2 = 0.520\,590$. 利用正态近似, 信度区间为 $2.499\,416 \pm 1.96(0.520\,590)^{1/2}$, 得到 $a = 1.085\,2$ 和 $b = 3.913\,6$. 这个区间 (在正态分布下) 是 HPD 的, 因为正态分布是对称的.

这个近似以后验众数 2.291 132 为中心 (见例 12.33). 后验分布密度 [由 (12.4)] 的负对数的二阶导数为

$$-\frac{\mathrm{d}^2}{\mathrm{d}\alpha^2}\ln\left[\frac{\alpha^{11}\mathrm{e}^{-4.801\ 121\alpha}}{(11!)(1/4.801\ 121)^{12}}\right] = \frac{11}{\alpha^2}.$$

取倒数得到方差的估计. 在 α 的众数估计点进行计算, 得到 $(2.291\ 132)^2/11 = 0.477\,208$, 由此信度区间为 $2.291\,13 \pm 1.96(0.477\,208)^{1/2}$, $a = 0.937\,2$ 和 $b = 3.645\,1$.□

预测分布也使用了相同的概念. 但是, 将不使用贝叶斯中心极限定理, 因为只需要预测一个值. 唯一可能的应用是, 对于较大容量的原始样本, 可以用多元正态分布来代替 (12.3) 中的真实后验分布.

例 12.41 (续例 12.28) 构造下一个观测的 95% 最高密度预测区间.

解 容易看出预测密度函数 (12.5) 是严格减的. 因此最高密度区域从 $a = 100$ 开始到 b. b 的值如下确定

$$\begin{aligned}0.95 &= \int_{100}^{b} \frac{12(4.801\ 121)^{12}}{y(0.195\ 951 + \ln y)^{13}}\mathrm{d}y \\ &= \int_{0}^{\ln(b/100)} \frac{12(4.801\ 121)^{12}}{(4.801\ 121 + x)^{13}}\mathrm{d}x \\ &= 1 - \left[\frac{4.801\ 121}{4.801\ 121 + \ln(b/100)}\right]^{12},\end{aligned}$$

解得 $b = 390.184\,0$. 有趣的是, 预测分布的众数为 100(因为概率密度函数是严格递减的), 而均值为无穷 ($b = \infty$ 和附加的 y 在被积函数中, 经过变形, 被积函数的形式如 $\mathrm{e}^x x^{-13}$, 当 x 趋向无穷的时候, 函数值趋于无穷). □

下面的例子重新回顾了 4.6.3 节中的计算. 在 4.6.3 节我们得知负二项分布为 Poisson 变量经 gamma 混合的结果. 下面的例子说明在贝叶斯背景下如何处理相同的计算.

例 12.42 已知保单每年的索赔数服从 Poisson 分布, 参数未知, 但是先验分布为 gamma 分布, 参数为 α 和 θ. 假设过去的一年该保单发生了 x 次索赔, 利用贝叶斯方法估计下一年的索赔数量. 假设过去 n 年发生的索赔数为 $x_1, \cdots, x_n$, 重新计算下一年的索赔数量.

解 关键的分布为 (这里 $x = 0, 1, \cdots, \lambda, \alpha, \theta > 0$)

先验: $\pi(\lambda) = \dfrac{\lambda^{\alpha-1}\mathrm{e}^{-\lambda/\theta}}{\Gamma(\alpha)\theta^\alpha}$.

模型：$p(x|\lambda) = \dfrac{\lambda^x e^{-\lambda}}{x!}$.

联合：$p(x,\lambda) = \dfrac{\lambda^{x+\alpha-1} e^{-(1+1/\theta)\lambda}}{x!\Gamma(\alpha)\theta^\alpha}$.

边缘：
$$p(x) = \int_0^\infty \frac{\lambda^{x+\alpha-1} e^{-(1+1/\theta)\lambda}}{x!\Gamma(\alpha)\theta^\alpha} d\lambda$$
$$= \frac{\Gamma(x+\alpha)}{x!\Gamma(\alpha)\theta^\alpha(1+1/\theta)^{x+\alpha}}$$
$$= \binom{x+\alpha-1}{x} \left(\frac{1}{1+\theta}\right)^\alpha \left(\frac{\theta}{1+\theta}\right)^x.$$

后验：
$$\pi(\lambda|x) = \frac{\lambda^{x+\alpha-1} e^{-(1+1/\theta)\lambda}}{x!\Gamma(\alpha)\theta^\alpha} \Big/ \frac{\Gamma(x+\alpha)}{x!\Gamma(\alpha)\theta^\alpha(1+1/\theta)^{x+\alpha}}$$
$$= \frac{\lambda^{x+\alpha-1} e^{-(1+1/\theta)\lambda}(1+1/\theta)^{x+\alpha}}{\Gamma(x+\alpha)}.$$

边缘分布是负二项分布，其中 $r=\alpha$, $\beta=\theta$. 后验分布是 gamma 分布，形状参数 "α" 等于 $x+\alpha$, 尺度参数 "θ" 等于 $(1+1/\theta)^{-1} = \theta/(1+\theta)$. Poisson 参数的贝叶斯估计为后验均值 $(x+\alpha)\theta/(1+\theta)$. (12.3) 式给出了预测分布,

$$\begin{aligned} p(y|x) &= \int_0^\infty \frac{\lambda^y e^{-\lambda}}{y!} \frac{\lambda^{x+\alpha-1} e^{-(1+1/\theta)\lambda}(1+1/\theta)^{x+\alpha}}{\Gamma(x+\alpha)} d\lambda \\ &= \frac{(1+1/\theta)^{x+\alpha}}{y!\Gamma(x+\alpha)} \int_0^\infty \lambda^{y+x+\alpha-1} e^{-(2+1/\theta)\lambda} d\lambda \\ &= \frac{(1+1/\theta)^{x+\alpha}\Gamma(y+x+\alpha)}{y!\Gamma(x+\alpha)(2+1/\theta)^{y+x+\alpha}}, \quad y=0,1,\cdots, \end{aligned}$$

重新整理后可以看出这是一个参数为 $r=x+\alpha$, $\beta=\theta/(1+\theta)$ 的负二项分布，所以下一年索赔数的期望为 $(x+\alpha)\theta/(1+\theta)$. 或者，由 (12.6) 式得到

$$E(Y|x) = \int_0^\infty \lambda \frac{\lambda^{x+\alpha-1} e^{-(1+1/\theta)\lambda}(1+1/\theta)^{x+\alpha}}{\Gamma(x+\alpha)} d\lambda = \frac{(x+\alpha)\theta}{1+\theta}.$$

对于容量为 n 的样本，关键点是现在的模型分布为

$$p(\boldsymbol{x}|\lambda) = \frac{\lambda^{x_1+\cdots+x_n} e^{-n\lambda}}{x_1!\cdots x_n!}.$$

沿着这个思路，后验分布仍然是 gamma 分布，形状参数为 $x_1+\cdots+x_n+\alpha = n\bar{x}+\alpha$, 尺度参数为 $\theta/(1+n\theta)$. 预测分布还是负二项分布，参数 $r=n\bar{x}+\alpha$, $\beta=\theta/(1+n\theta)$.□

在只需要矩的情况, 双重期望公式将变得非常有用. 只要矩存在, 对任意的随机变量 X 和 Y, 有

$$\mathrm{E}(Y) = \mathrm{E}[\mathrm{E}(Y|X)], \tag{12.8}$$

$$\mathrm{Var}(Y) = \mathrm{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}[\mathrm{E}(Y|X)]. \tag{12.9}$$

对于预测分布,

$$\begin{aligned}
\mathrm{E}(Y|\boldsymbol{x}) &= \mathrm{E}_{\Theta|\boldsymbol{x}}[\mathrm{E}(Y|\Theta,\boldsymbol{x})] = \mathrm{E}_{\Theta|\boldsymbol{x}}[\mathrm{E}(Y|\Theta)],\\
\mathrm{Var}(Y|\boldsymbol{x}) &= \mathrm{E}_{\Theta|\boldsymbol{x}}[\mathrm{Var}(Y|\Theta,\boldsymbol{x})] + \mathrm{Var}_{\Theta|\boldsymbol{x}}[\mathrm{E}(Y|\Theta,\boldsymbol{x})]\\
&= \mathrm{E}_{\Theta|\boldsymbol{x}}[\mathrm{Var}(Y|\Theta)] + \mathrm{Var}_{\Theta|\boldsymbol{x}}[\mathrm{E}(Y|\Theta)].
\end{aligned}$$

其中的条件期望值和方差是简化的, 因为当 Θ 已知时, $\boldsymbol{x}$ 并没有提供任何关于 Y 的附加信息. 这里只是将 (12.6) 式重新表述. □

例 12.43　*应用上述公式计算前例的预测均值和方差. 然后预测第 16 章的信度公式.*

解　预测均值为 $\mathrm{E}(Y|\lambda) = \lambda$, 则

$$\mathrm{E}(Y|\boldsymbol{x}) = \mathrm{E}(\lambda|\boldsymbol{x}) = \frac{(n\bar{x}+\alpha)\theta}{1+n\theta}.$$

预测方差为 $\mathrm{Var}(Y|\lambda) = \lambda$, 则

$$\begin{aligned}
\mathrm{Var}(Y|\boldsymbol{x}) &= \mathrm{E}(\lambda|\boldsymbol{x}) + \mathrm{Var}(\lambda|\boldsymbol{x})\\
&= \frac{(n\bar{x}+\alpha)\theta}{1+n\theta} + \frac{(n\bar{x}+\alpha)\theta^2}{(1+n\theta)^2}\\
&= (n\bar{x}+\alpha)\frac{\theta}{1+n\theta}\left(1+\frac{\theta}{1+n\theta}\right).
\end{aligned}$$

以上结果和已知 y 服从负二项分布时的均值和方差一致. 但是这些量是由模型 (Poisson) 和后验 (gamma) 分布的矩得到的. 预测的均值可以表示为

$$\frac{n\theta}{1+n\theta}\bar{x} + \frac{1}{1+n\theta}\alpha\theta,$$

它表示数据均值与先验分布均值的加权平均. 可以发现当样本容量增加时, 更多的权重放在数据上, 而先验分布的权重变小了. 先验分布的方差随着 θ 的增大会增加, 当然, 数据的权重也随之增大了. 第 16 章的一般信度公式可表示为基于数据的估计和先验看法的加权平均. □

12.4.3　共轭先验分布和线性指数族

有一个应用范围很广的参数分布族在贝叶斯分析方法中具有特殊的地位, 它包括了迄今为止我们遇到的很多分布. 其定义如下.

定义 12.44 称随机变量 X(离散或者连续的) 的分布属于**线性指数族**, 若它的概率函数可以用参数 θ 参数化表示为

$$f(x;\theta)=\frac{p(x)\mathrm{e}^{-\theta x}}{q(\theta)}. \tag{12.10}$$

其中函数 $p(x)$ 只依赖于 x(不依赖 θ), 函数 $q(\theta)$ 是正规化常数. 另外, 还必须要求随机变量的支集不依赖于 θ. 参数 θ 被称作分布的**自然参数**.

例 12.45 证明指数分布具有 (12.10) 式的形式.

解 概率密度函数为

$$f(x;\beta)=\beta^{-1}\mathrm{e}^{-\beta^{-1}x}.$$

如果令 $\theta=1/\beta$, 则概率密度函数为

$$f(x;\theta)=\frac{1\mathrm{e}^{-\theta x}}{\theta^{-1}},$$

即为 (12.10) 式的形式, $p(x)=1$ 和 $q(\theta)=1/\theta$. □

例 12.46 证明 Poisson 分布属于线性指数族.

解 概率函数为

$$f(x;\lambda)=\frac{\lambda^x\mathrm{e}^{-\lambda}}{x!}=\frac{(1/x!)\mathrm{e}^{-(-\ln\lambda)x}}{\mathrm{e}^{\lambda}}.$$

令 $\theta=-\ln\lambda$, 则概率函数为

$$f(x;\theta)=\frac{(1/x!)\mathrm{e}^{-\theta x}}{\mathrm{e}^{\mathrm{e}^{-\theta}}},$$

亦为 (12.10) 式的形式, 其中 $p(x)=1/x!$ 和 $q(\theta)=\mathrm{e}^{\mathrm{e}^{-\theta}}$. 注意, 在这个表示中 Poisson 分布的均值为 $\mathrm{e}^{-\theta}$. □

例 12.47 证明均值为 μ, 方差为 v 的正态分布属于线性指数族.

解 概率密度函数为

$$\begin{aligned}f(x;\mu,v)&=(2\pi v)^{-1/2}\exp\left[-\frac{1}{2v}(x-\mu)^2\right]\\&=(2\pi v)^{-1/2}\exp\left(-\frac{x^2}{2v}+\frac{\mu}{v}x-\frac{\mu^2}{2v}\right)\\&=\frac{(2\pi v)^{-1/2}\exp\left(-\frac{x^2}{2v}\right)\exp\left(\frac{\mu}{v}x\right)}{\exp\left(\frac{\mu^2}{2v}\right)}.\end{aligned}$$

令 $\theta=-\mu/v$, 则概率密度函数为

$$f(x;\theta,v)=\frac{(2\pi v)^{-1/2}\exp\left(-\frac{x^2}{2v}\right)\exp(-\theta x)}{\exp\left(\frac{\theta^2 v}{2}\right)},$$

即为 (12.10) 式的形式, 其中 $p(x) = (2\pi v)^{-1/2}\exp[-x^2/(2v)]$, $q(\theta) = \exp(\theta^2 v/2)$. □

现在计算由 (12.10) 定义的分布的均值和方差. 首先, 注意到

$$\ln f(x;\theta) = \ln p(x) - \theta x - \ln q(\theta).$$

对 θ 求微分得到

$$\frac{\partial}{\partial\theta} f(x;\theta) = \left[-x - \frac{q'(\theta)}{q(\theta)}\right] f(x;\theta). \tag{12.11}$$

在 x(已知不依赖 θ) 的定义域上积分 (或求和) 得到

$$\int \frac{\partial}{\partial\theta} f(x;\theta)\mathrm{d}x = -\int x f(x;\theta)\mathrm{d}x - \frac{q'(\theta)}{q(\theta)}\int f(x;\theta)\mathrm{d}x.$$

在上式左端交换积分 (或求和) 和微分的顺序得到

$$\frac{\partial}{\partial\theta}\left[\int f(x;\theta)\mathrm{d}x\right] = -\int x f(x;\theta)\mathrm{d}x - \frac{q'(\theta)}{q(\theta)}\int f(x;\theta)\mathrm{d}x.$$

由已知 $\int f(x;\theta)\mathrm{d}x = 1$ 和 $\int x f(x;\theta)\mathrm{d}x = \mathrm{E}(X)$, 有

$$\frac{\partial}{\partial\theta}(1) = -\mathrm{E}(X) - \frac{q'(\theta)}{q(\theta)}.$$

即均值为

$$\mathrm{E}(X) = \mu(\theta) = -\frac{q'(\theta)}{q(\theta)} = -\frac{\mathrm{d}}{\mathrm{d}\theta}\ln q(\theta). \tag{12.12}$$

为了得到方差, (12.11) 式可以改写为

$$\frac{\partial}{\partial\theta} f(x;\theta) = -[x-\mu(\theta)] f(x;\theta).$$

再对 θ 求导得到

$$\begin{aligned}\frac{\partial^2}{\partial\theta^2} f(x;\theta) &= \mu'(\theta) f(x;\theta) - [x-\mu(\theta)]\frac{\partial}{\partial\theta} f(x;\theta)\\ &= \mu'(\theta) f(x;\theta) + [x-\mu(\theta)]^2 f(x;\theta).\end{aligned}$$

在 x 的定义域上求积分得到

$$\int \frac{\partial^2}{\partial\theta^2} f(x,\theta)\mathrm{d}x = \mu'(\theta)\int f(x;\theta)\mathrm{d}x + \int [x-\mu(\theta)]^2 f(x;\theta)\mathrm{d}x.$$

也就是

$$\int [x-\mu(\theta)]^2 f(x;\theta)\mathrm{d}x = -\mu'(\theta) + \frac{\partial^2}{\partial\theta^2}\int f(x;\theta)\mathrm{d}x.$$

因为 $\mu(\theta)$ 为均值, 左边是方差 (由定义知), 因右边第二项为零, 得到

$$\mathrm{Var}(X) = v(\theta) = -\mu'(\theta) = \frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \ln q(\theta). \tag{12.13}$$

在例 12.42 中, 后验分布与先验分布 (gamma) 的类型相同. 这使得计算相对简单, 类似这个概念的定义如下.

定义 12.48 称先验分布为给定模型的共轭先验分布, 如果其后验分布与先验分布具有相同的类型 (只是参数不同).

下面的定理证明了如果模型为线性指数族, 则很容易得到共轭先验分布.

定理 12.49 已知给定 $\Theta=\theta$ 时, 随机变量 $X_1, \cdots, X_n$ 是独立同分布的, 其概率函数为

$$f_{X_j|\Theta}(X_j|\theta) = \frac{p(x_j)\mathrm{e}^{-\theta x_j}}{q(\theta)},$$

其中 Θ 的概率密度函数为

$$\pi(\theta) = \frac{[q(\theta)]^{-k}\mathrm{e}^{-\theta\mu k}}{c(\mu, k)},$$

其中 k 和 μ 是分布的参数, $c(\mu,k)$ 是正规化常数. 则后验分布概率函数 $\pi_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$ 与 $\pi(\theta)$ 具有相同的形式.

证明 后验分布为

$$\begin{aligned}
\pi(\theta|\mathbf{x}) &\propto \frac{\left[\prod_{j=1}^{n} p(x_j)\right]\mathrm{e}^{-\theta\Sigma x_j}}{q(\theta)^n} \frac{[q(\theta)^{-k}\mathrm{e}^{-\theta\mu k}]}{c(\mu,k)} \\
&\propto [q(\theta)]^{-(k+n)} \exp\left[\left(-\theta\frac{\mu k + \sum x_j}{k+n}\right)(k+n)\right] \\
&\propto [q(\theta)]^{-k^*} \exp(-\theta\mu^* k^*),
\end{aligned}$$

与 $\pi(\theta)$ 具有相同的形式, 参数为

$$\begin{aligned}
k^* &= k+n, \\
\mu^* &= \frac{\mu k + \sum x_j}{k+n} = \frac{k}{k+n}\mu + \frac{n}{k+n}\bar{x}.
\end{aligned}$$

□

例 12.50 证明对于 Poisson 模型, 定理 12.49 中的共轭先验分布是 gamma 分布.

解 由例 12.46, 有 $q(\theta)=\exp(\mathrm{e}^{-\theta})$ 和 $\lambda=\exp(-\theta)$. 定理给出的先验分布为

$$\pi(\theta) \propto [\exp(\mathrm{e}^{-\theta})]^{-k}\exp(-\theta\mu k).$$

则 λ 的先验密度为

$$\pi(\lambda) \propto [\exp(\lambda)]^{-k}\lambda^{\mu k}\lambda^{-1} = \lambda^{\mu k-1}\mathrm{e}^{-\lambda k},$$

这是一个 gamma 分布, 其中 $\alpha = \mu k$ 和 $\theta = 1/k$, λ^{-1} 项是在变量替换中出现的, 等于 $|\mathrm{d}\theta/\mathrm{d}\lambda|$. □

12.4.4　计算问题

到目前可以明显看出, 贝叶斯分析是通过引入积分或者求和进行的, 所以至少在理论上贝叶斯分析是可行的. 但是, 只有在极少的情况下这些积分或求和是易于计算的, 这就意味着大多数贝叶斯分析需要数值积分. 尽管一维的积分容易得到较高精度的结果, 但高维积分的估计就要困难得多. 为了解决这个问题, 已经进行了很多的努力, 并且也得到了许多有独创性的方法. 其中的一部分在 Klugman[77] 中进行了概括性的介绍. 但现在广泛使用的方法为 Markov 链 Monte Carlo 模拟. [118] 提供了这个方法的详细讨论, 其精算应用在 [21] 和 [119] 中有讨论.

还有一种方法可以完全避免计算问题. 下面的例子 (使用简化的形式) 对此进行了说明, 这个例子来自于 Meyers[92], 它也使用了这种技术. 这个例子还说明了如何使用贝叶斯分析估计参数的函数.

例 12.51　现有 100 个超过 100 000 的损失数据. 使用单参数 Pareto 分布建模, 其中 $\theta = 100\ 000$, α 未知. 目标是估计 1 000 000~5 000 000 这个区间的平均损失. 由观测, 有 $\sum\limits_{j=1}^{100} \ln x_j = 1\ 208.435\ 4$.

解　模型的密度函数为

$$\begin{aligned} f_{\boldsymbol{X}|A}(\boldsymbol{x}|\alpha) &= \prod_{j=1}^{100} \frac{\alpha(100\ 000)^{\alpha}}{x_j^{\alpha+1}} \\ &= \exp\left[100\ln\alpha + 100\alpha\ln 100\ 000 - (\alpha+1)\sum_{j=1}^{100}\ln x_j\right] \\ &= \exp\left(100\ln\alpha - \frac{100\alpha}{1.75} - 1\ 208.435\ 4\right). \end{aligned}$$

这个密度出现在表 12-6 的第 3 列. 为了防止计算溢出, 没有在求幂之前减去 1 208.435 4. 这使这些值和真实的密度函数成比例. 在第 2 列中给出先验密度, 选择的根据是认为真值在 1~2.5 这个范围内, 并且相对端点来说真值更可能在 1.5 附近. 后验密度由 (12.2) 得到, 第 4 列中给出了这些分子. 分布函数不再是积分, 而是求和. 第 4 列的最后给出了求和值, 比例变换后的结果列在第 5 列.

表 12-6 区间内平均损失的贝叶斯估计

α	$\pi(\alpha)$	$f(\boldsymbol{x}\|\alpha)$	$\pi(\alpha)f(\boldsymbol{x}\|\alpha)$	$\pi(\alpha\|\boldsymbol{x})$	LAS(α)	$\pi\times$L*	$\pi(\alpha\|\boldsymbol{x})l(\alpha)^2$
1.0	0.040 0	1.52×10^{-25}	6.10×10^{-27}	0.000 0	160 944	0	6 433
1.1	0.049 6	6.93×10^{-24}	3.44×10^{-25}	0.000 0	118 085	2	195 201
1.2	0.059 2	1.37×10^{-22}	8.13×10^{-24}	0.000 3	86 826	29	2 496 935
1.3	0.068 8	1.36×10^{-21}	9.33×10^{-23}	0.003 8	63 979	243	15 558 906
1.4	0.078 4	7.40×10^{-21}	5.80×10^{-22}	0.023 6	47 245	1 116	52 737 840
1.5	0.088 0	2.42×10^{-20}	2.13×10^{-21}	0.086 7	34 961	3 033	106 021 739
1.6	0.083 2	5.07×10^{-20}	4.22×10^{-21}	0.171 8	25 926	4 454	115 480 050
1.7	0.078 4	7.18×10^{-20}	5.63×10^{-21}	0.229 3	19 265	4 418	85 110 453
1.8	0.073 6	7.19×10^{-20}	5.29×10^{-21}	0.215 6	14 344	3 093	44 366 353
1.9	0.068 8	5.29×10^{-20}	3.64×10^{-21}	0.148 2	10 702	1 586	16 972 802
2.0	0.064 0	2.95×10^{-20}	1.89×10^{-21}	0.076 8	8 000	614	4 915 383
2.1	0.059 2	1.28×10^{-20}	7.57×10^{-22}	0.030 8	5 992	185	1 106 259
2.2	0.054 4	4.42×10^{-21}	2.40×10^{-22}	0.009 8	4 496	44	197 840
2.3	0.049 6	1.24×10^{-21}	6.16×10^{-23}	0.002 5	3 380	8	28 650
2.4	0.044 8	2.89×10^{-22}	1.29×10^{-23}	0.000 5	2 545	1	3 413
2.5	0.040 0	5.65×10^{-23}	2.26×10^{-24}	0.000 1	1 920	0	339
	1.000 0		2.46×10^{-20}	1.000 0		18 827	445 198 597

* $\pi(\alpha|\boldsymbol{x})LAS(\alpha)$

从第 5 列可以看出, 后验分布的众数为 $\alpha=1.7$, 而最大似然估计为 1.75(见习题 12.69). 后验分布 α 的均值可以通过对第 1 列和第 5 列的乘积求和得到. 这里我们更关心区间内的平均损失. 具体为

$$\begin{aligned}\text{LAS}(\alpha)&=\text{E}(X\wedge 5\ 000\ 000)-\text{E}(X\wedge 1\ 000\ 000)\\&=\begin{cases}\dfrac{100\ 000^{\alpha}}{\alpha-1}\left(\dfrac{1}{1\ 000\ 000^{\alpha-1}}-\dfrac{1}{5\ 000\ 000^{\alpha-1}}\right), & \alpha\neq 1,\\ 100\ 000(\ln 5\ 000\ 000-\ln 1\ 000\ 000), & \alpha=1.\end{cases}\end{aligned}$$

第 6 列为 LAS(α) 的 16 个可能值. 最后 2 列是为了得到后验分布的区间平均损失的期望值. 点估计是后验的均值 18 827, 后验标准差为

$$\sqrt{445\ 198\ 597-18\ 827^2}=9\ 526.$$

我们还可以用第 5 列和第 6 列构造一个信度区间. 去掉前 5 行和后 4 行共减少了 0.040 6 的后验概率. 这样剩下的 (5 992, 34 961) 构成了一个区间平均损失 96% 的信度区间. Meryes 认为即便样本量相当大, 估计的准确性还是相当糟糕.

后验分布的离散估计的精度可以通过观测更多的值来提高. 这几乎没有增加处理电子数据表的难度. 这里使用小样本只是为了说明方法. □

习题

12.67 证明: 若 Y 服从例 12.28 的预测分布, 则 $\ln Y - \ln 100$ 服从 Pareto 分布.

12.68 试计算例 12.28 中 α 的后验分布, 假设先验分布为任意的 gamma 分布, 为了避免混淆, gamma 分布的第一个参数用 γ 表示. 然后确定 gamma 参数的一个特定组合, 使得后验均值是 α 的最大似然估计值, 而不依赖于 $x_1, \cdots, x_n$ 的特定取值. 这个先验分布是非正常的吗?

12.69 证明例 12.51 中 α 的最大似然估计值为 1.75.

12.70 令 $x_1, \cdots, x_n$ 为来自于参数 μ 和 σ 未知的对数正态分布的随机样本. 令先验密度为 $\pi(\mu, \sigma) = \sigma^{-1}$.

(a) 给出 μ 和 σ 的后验概率密度函数的形式.

(b) 利用后验众数确定 μ 和 σ 的贝叶斯估计.

(c) 固定 σ 为 (b) 中众数确定的值, 然后确定 μ 的准确 (条件) 概率密度函数. 然后利用这个密度函数确定 μ 的 95% 的 HPD 信度区间.

12.71 已知来自 gamma 分布的 100 个随机样本, 并已知 α 为 2、θ 未知, 有 $\sum\limits_{j=1}^{100} x_j = 30\ 000$. θ 的先验分布为逆 gamma, 其参数 α 用 β 代替, θ 用 λ 代替.

(a) 确定 θ 的准确后验分布. 此外, β 和 λ 的值未知.

(b) 总体的均值为 2θ. 确定 2θ 的后验均值, 首先令先验分布的参数为 $\beta = \lambda = 0$[等价于 $\pi(\theta) = \theta^{-1}$], 然后令 $\beta = 2$ 和 $\lambda = 250$(先验均值为 250). 然后, 对于每种情况, 确定两端分别有 2.5% 概率的 95% 的信度区间.

(c) 确定 2θ 的后验方差, 然后对于 (b) 中的两个先验分布, 分别利用贝叶斯中心极限定理构造 95% 的信度区间.

(d) 确定 θ 的最大似然估计, 然后利用估计方差构造 2θ 的 95% 置信区间.

12.72 已知给定 $\Theta = \theta$ 时随机变量 $X_1, \cdots, X_n$ 为独立的且均服从二项分布, 其概率函数为

$$f_{X_j|\Theta}(x_j|\theta) = \binom{K_j}{x_j} \theta^{x_j}(1-\theta)^{K_j - x_j}, \quad x_j = 0, 1, \cdots, K_j,$$

Θ 本身服从参数为 a 和 b 的贝塔分布, 其概率密度函数为

$$\pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}, \quad 0 < \theta < 1.$$

(a) 验证 X_j 的边缘概率函数为

$$f_{X_j}(x_j) = \frac{\binom{-a}{x_j}\binom{-b}{K_j - x_j}}{\binom{-a-b}{K_j}}, \quad x_j = 0, 1, \cdots, K_j,$$

并且 $\mathrm{E}(X_j) = aK_j/(a+b)$. 这个分布被称为二项贝塔或者负超几何分布.

(b) 确定后验概率密度函数 $\pi_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$ 和后验均值 $\mathrm{E}(\Theta|\boldsymbol{x})$.

12.73 已知给定 $\Theta=\theta$ 时随机变量 $X_1,\cdots,X_n$ 为独立的且服从同一个指数分布, 其概率密度函数为

$$f_{X_j|\Theta}(x_j|\theta)=\theta \mathrm{e}^{-\theta x_j},\quad x_j>0,$$

Θ 本身服从参数为 $\alpha>1$ 和 $\beta>0$ 的 gamma 分布,

$$\pi(\theta)=\frac{\theta^{\alpha-1}\mathrm{e}^{-\theta/\beta}}{\Gamma(\alpha)\beta^{\alpha}},\quad \theta>0.$$

(a) 验证 X_j 的边缘概率密度函数为

$$f_{X_j}(x_j)=\alpha\beta^{-\alpha}(\beta^{-1}+x_j)^{-\alpha-1},\quad x_j>0,$$

$$\mathrm{E}(X_j)=\frac{1}{\beta(\alpha-1)}.$$

这是某种形式的 Pareto 分布.

(b) 确定后验概率密度函数 $\pi_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$ 和后验均值 $E(\Theta|\boldsymbol{x})$.

12.74 已知给定 $\Theta=\theta$ 时随机变量 $X_1,\cdots,X_n$ 独立且服从参数为 r 和 θ 的负二项分布, 其概率函数为

$$f_{X_j|\Theta}(x_j|\theta)=\begin{pmatrix} r+x_j-1 \\ x_j \end{pmatrix}\theta^{r}(1-\theta)^{x_j},\quad x_j=0,1,2,\cdots,$$

Θ 本身服从参数为 a 和 b 的 beta 分布, 概率密度函数为

$$\pi(\theta)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1},\quad 0<\theta<1.$$

(a) 验证 X_j 的边缘概率函数为

$$f_{X_j}(x_j)=\frac{\Gamma(r+x_j)}{\Gamma(r)x_j!}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+r)\Gamma(b+x_j)}{\Gamma(a+r+b+x_j)},\quad x_j=0,1,2,\cdots,$$

而且有

$$\mathrm{E}(X_j)=\frac{rb}{a-1}.$$

这个分布被称作**一般 Waring 分布**.$b=1$ 的特例是 **Waring 分布**, 而 $r=1$ 和 $b=1$ 时称为 **Yule 分布**.

(b) 确定后验概率密度函数 $f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$ 和后验均值 $\mathrm{E}(\Theta|\boldsymbol{x})$.

12.75 已知给定 $\Theta=\theta$ 时随机变量 $X_1,\cdots,X_n$ 独立, 且服从相同的正态分布, 均值为 μ, 方差为 θ^{-1}, Θ 服从参数为 α 和 (θ 替换为)$1/\beta$ 的 gamma 分布.

(a) 验证 X_j 的边缘概率密度函数为

$$f_{X_j}(x_j)=\frac{\Gamma\left(\alpha+\frac{1}{2}\right)}{\sqrt{2\pi\beta}\Gamma(\alpha)}\left[1+\frac{1}{2\beta}(x_j-\mu)^2\right]^{-\alpha-1/2},\quad -\infty<x_j<\infty,$$

这是 t- 分布.

(b) 确定后验概率密度函数 $f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$ 和后验均值 $\mathrm{E}(\Theta|\boldsymbol{x})$.

12.76　证明对于具有如下概率函数的二项分布

$$f(x;p)=\binom{n}{x}p^x(1-p)^{n-x},$$

属于 (12.10) 式的分布族, 并确定 $\theta, p(x)$ 和 $q(\theta)$.

12.77　考虑负二项分布, 概率函数为

$$f(x;\alpha,\beta)=\frac{\Gamma(\alpha+x)}{\Gamma(\alpha)x!}\left(\frac{\beta}{1+\beta}\right)^{\alpha}\left(\frac{1}{1+\beta}\right)^{x}.$$

如果 α 是确定的, 证明 $f(x;\alpha,\beta)$ 具有 (12.10) 形式, 并确定 $\theta, p(x)$ 和 $q(\theta)$.

12.78　假设 $X_1,\cdots,X_n$ 是独立同分布的, 分布的形式如 (12.10) 式. 证明均值的最大似然估计是样本均值. 也就是说, 如果 $\hat{\theta}$ 是 θ 的最大似然估计量, 证明

$$\widehat{\mu(\theta)}=\mu(\hat{\theta})=\overline{X}.$$

12.79　考虑 (12.10) 式如下给出的一般化形式

$$f(x;\theta)=\frac{p(m,x)\mathrm{e}^{-m\theta x}}{[q(\theta)]^m},$$

其中 m 为已知参数. 证明均值仍然由 (12.12) 式给出但是方差变为 $v(\theta)/m$, 其中 $v(\theta)$ 由 (12.13) 式给出.

12.80　设 $X_1,\cdots,X_n$ 独立同分布, 关于 Θ 的条件概率函数为

$$f_{X_j|\Theta}(x_j|\theta)=\frac{p(x_j)\mathrm{e}^{-\theta x_j}}{q(\theta)}.$$

令 $S=X_1+\cdots+X_n$.

(a) 证明, S 关于 Θ 的条件概率函数形式为

$$f_{S|\Theta}(s|\theta)=\frac{p_n(s)\mathrm{e}^{-\theta s}}{[q(\theta)]^n},$$

其中 $p_n(s)$ 不依赖 θ.

(b) 证明后验分布 $\pi_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$ 与 $\Theta|S$ 的 (条件) 分布相同,

$$\pi_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})=\frac{f_{S|\Theta}(s|\theta)\pi(\theta)}{f_S(s)},$$

其中 $\pi(\theta)$ 是 Θ 的概率函数, 而 $f_S(s)$ 是 S 的边缘概率函数.

12.81　当给定 N 时随机变量 X 服从参数为 N 和 p 的二项分布.

(a) 证明, 如果 N 为 Poisson 分布, 则 X 相同 (无条件的), 并确定参数.

(b) 证明, 如果 N 为二项分布, 则 X 相同 (无条件的), 并确定参数.

(c) 证明, 如果 N 为负二项分布, 则 X 相同 (无条件的), 并确定参数.

12.82*　从装有 2 个 6 面骰子的瓮中随机选取一个投掷, 第 1 个骰子的 3 个面为数字 2 其他 3 个面为 1,3, 4. 第 2 个骰子有 3 个面为数字 4 其他的 3 个面分别为 1,2,3. 前 5 次顺序投掷的数字为 2,3,4,1,4. 试计算任选的骰子为第 2 个的概率.

12.83* 一年中的索赔次数 Y 服从含参数 θ 的分布. 作为一个随机变量, Θ 服从 $(0,1)$ 区间上的均匀分布. Y 取 0 的无条件概率大于 0.35. 对于下面每一个条件概率函数, 确定其是否可能为 Y 的真实条件概率函数.

(a) $\Pr(Y=y|\theta)=\mathrm{e}^{-\theta}\theta^y/y!$.

(b) $\Pr(Y=y|\theta)=(y+1)\theta^2(1-\theta)^y$.

(c) $\Pr(Y=y|\theta)=\begin{pmatrix}2\\y\end{pmatrix}\theta^y(1-\theta)^{2-y}$.

12.84* 关于未知值 H, 你的先验分布为 $\Pr\left(H=\frac{1}{4}\right)=\frac{4}{5}$ 和 $\Pr\left(H=\frac{1}{2}\right)=\frac{1}{5}$. 单次试验的观测服从分布 $\Pr(D=d|H=h)=h^d(1-h)^{1-d}$, $d=0,1$. 现有一次试验的结果为 $d=1$, 试给出 H 的后验分布.

12.85* 一年中的索赔次数 Y 服从参数为 θ 的 Poisson 分布. 参数 θ 服从指数分布, 概率密度函数为 $\pi(\theta)=\mathrm{e}^{-\theta}$. 某被保险人在一年中没有索赔. 对于这个被保险人, 确定 θ 的后验分布.

12.86* 一年中的索赔次数 Y 服从参数为 θ 的 Poisson 分布. 先验分布服从 gamma 分布, 其概率密度函数为 $\pi(\theta)=\theta\mathrm{e}^{-\theta}$. 已知在一年中有一次索赔, 确定 θ 的后验概率密度函数.

12.87* 已知每辆汽车的索赔数服从参数为 λ 的 Poisson 分布, 所有汽车有相同的参数. 先验分布为 gamma 分布, 参数 $\alpha=50$ 和 $\theta=1/500$. 在 2 年中, 保险人分别在第 1 年和第 2 年承保了 750 辆和 1100 辆汽车. 第 1 年和第 2 年分别有 65 和 112 个索赔. 确定后验 gamma 分布的变异系数.

12.88* 某个体一年内的索赔次数 r 服从二项分布, 概率函数为 $f(r)=\begin{pmatrix}3\\r\end{pmatrix}\theta^r(1-\theta)^{3-r}$. θ 的先验分布的概率密度函数为 $\pi(\theta)=6(\theta-\theta^2)$. 已知在一年中有一个索赔, 确定 θ 的后验概率密度函数.

12.89* 某个体一年中的索赔次数服从参数为 λ 的 Poisson 分布. λ 的先验分布为均值 0.14 和方差 0.000 4 的 gamma 分布. 过去的两年中观测到 110 个索赔. 已知每年有 310 个有效保单. 确定 λ 的后验分布的期望值和方差.

12.90* 某个体一年中的索赔次数服从参数为 λ 的 Poisson 分布. λ 的先验分布服从期望值为 2 的指数分布. 在第一年中有 3 个索赔. 确定 λ 的后验分布.

12.91* 一年中的索赔次数服从 $n=3$ 和 θ 未知的二项分布. θ 先验分布为贝塔分布, 概率密度函数为 $\pi(\theta)=280\theta^3(1-\theta)^4, 0<\theta<1$. 观测到两个索赔, 分别计算:

(a) θ 的后验分布;

(b) 从后验分布得到的 θ 的期望值.

12.92* 某个体风险为每年恰有一个索赔. 索赔量的概率密度函数为 $f(x)=t\mathrm{e}^{-tx}, x>0$. 参数 t 的先验分布为 $\pi(t)=t\mathrm{e}^{-t}$. 现观测索赔量 5, 试确定 t 的后验概率密度函数.

12.93 已知给定 $\Theta_1=\theta_1$ 和 $\Theta_2=\theta_2$ 时随机变量 $X_1,\cdots,X_n$ 独立且服从同一个均值为 θ_1, 方差为 θ_2^{-1} 的正态分布. 另外假设给定 $\Theta_2=\theta_2$ 时 Θ_1 的条件分布是一个均值为 μ, 方差为 σ^2/θ_2 的正态分布, 且 Θ_2 服从参数为 α 和 $\theta=1/\beta$ 的 gamma 分布.

(a) 证明, 给定 $\Theta_2=\theta_2$ 时, Θ_1 的后验条件分布为正态分布, 均值为

$$\mu_*=\frac{1}{1+n\sigma^2}\mu+\frac{n\sigma^2}{1+n\sigma^2}\bar{x},$$

方差为

$$\sigma_*^2=\frac{\sigma^2}{\theta_2(1+n\sigma^2)},$$

并且 Θ_2 的后验边缘分布为 gamma 分布, 参数为

$$\alpha_*=\alpha+\frac{n}{2},\quad \beta_*=\beta+\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2+\frac{n(\bar{x}-\mu)^2}{2(1+n\sigma^2)}.$$

(b) 计算后验边缘分布的均值 $\mathrm{E}(\Theta_1|x)$ 和 $\mathrm{E}(\Theta_2|x)$.

12.5 离散分布的估计

12.5.1 Poisson 分布

本章前面讨论的关于连续模型估计的基本原理同样可以应用于频率分布, 以 Poisson 模型为例说明.

例 12.52 表 12-7 给出了某医疗责任保单 10 年的索赔数. 使用矩方法和最大似然法估计 Poisson 参数.

表 12-7 医疗责任险年索赔数

年数	索赔数
1985	6
1986	2
1987	3
1988	0
1989	2
1990	1
1991	2
1992	5
1993	1
1994	3

解 可以按照不同的方法来对这些数据进行综合. 可以统计索赔数发生的年数, 例如零索赔的年数、一个索赔的年数等, 如表 12-8.

在 1985—1994 间的索赔总数为 25. 因此, 平均每年的索赔次数为 2.5, 这个平均值也可由表 12-8 算出. 令 n_k 表示恰好出现 k 次索赔的年数. 期望频率 (样本均值) 为

表 12-8 医疗责任险索赔频数

频数 (k)	观测数 (n_k)
0	1
1	2
2	3
3	2
4	0
5	1
6	1
7+	0

$$\overline{x}=\frac{\sum_{k=0}^{\infty}kn_k}{\sum_{k=0}^{\infty}n_k},$$

其中 n_k 表示频率 k 的观测数. 因此矩方法的 Poisson 参数估计为 $\hat{\lambda}=2.5$.

用这些数据很容易得到最大似然估计. 在观测值 k 的似然函数为 p_k, 则整个观测集的似然函数为

$$L=\prod_{k=0}^{\infty}p_k^{n_k}.$$

对数似然函数为

$$l=\sum_{k=0}^{\infty}n_k\ln p_k.$$

似然函数和对数似然函数均为未知参数的函数. Poisson 分布只有一个参数, 最大值计算比较容易.

对于 Poisson 分布, 有

$$p_k=\frac{\mathrm{e}^{-\lambda}\lambda^k}{k!},\quad \ln p_k=-\lambda+k\ln\lambda-\ln k!.$$

对数似然函数为

$$l=\sum_{k=0}^{\infty}n_k(-\lambda+k\ln\lambda-\ln k!)=-\lambda n+\sum_{k=0}^{\infty}kn_k\ln\lambda-\sum_{k=0}^{\infty}n_k\ln k!,$$

其中 $n=\sum\limits_{k=0}^{\infty}n_k$ 为样本容量. 对数似然函数关于 λ 求微分, 得到

$$\frac{\mathrm{d}l}{\mathrm{d}\lambda}=-n+\sum_{k=0}^{\infty}kn_k\frac{1}{\lambda}.$$

令对数似然函数的导数值为零, 求解后可以得到最大似然估计量. 估计量为

$$\hat{\lambda}=\frac{\sum_{k=0}^{\infty}kn_k}{n}=\overline{x}.$$

由此可见对于 Poisson 分布最大似然估计量和矩方法估计量是一样的.

如果 N 服从均值为 λ 的 Poisson 分布, 则

$$\mathrm{E}(\hat{\lambda}) = \mathrm{E}(N) = \lambda, \quad \mathrm{Var}(\hat{\lambda}) = \frac{\mathrm{Var}(N)}{n} = \frac{\lambda}{n}.$$

因此, $\hat{\lambda}$ 是无偏的相合估计. 由定理 12.13, 最大似然估计量渐近服从正态分布, 均值为 λ, 方差为

$$\begin{aligned}
\mathrm{Var}(\hat{\lambda}) &= \left\{-n\mathrm{E}\left[\frac{d^2}{d\lambda^2}\ln p_N\right]\right\}^{-1} \\
&= \left\{-n\mathrm{E}\left[\frac{d^2}{d\lambda^2}(-\lambda + N\ln\lambda - \ln N!)\right]\right\}^{-1} \\
&= [n\mathrm{E}(N/\lambda^2)]^{-1} \\
&= (n\lambda^{-1})^{-1} = \frac{\lambda}{n}.
\end{aligned}$$

在这种情况下方差的渐近估计值等于它的真值. 由这个信息, 可以构造参数真值的近似 95% 置信区间 $\hat{\lambda} \pm 1.96(\hat{\lambda}/n)^{1/2}$. 本例的区间为 $(1.52, 3.48)$, 因为是基于大样本理论得到的这个置信区间因此只是近似. 本例的样本容量非常小, 应该小心使用置信区间. □

至今出现的公式都假设在每个观测频率的观测次数都有记录. 有时数据并没有给出这些信息, 最常见的例子就是给出 $k+$ 的最后记录, 它表示 k 次以上的索赔观测. 如果 n_{k+} 表示这样的观测次数, 则似然函数为

$$(p_k + p_{k+1} + \cdots)^{n_{k+}} = (1 - p_0 - \cdots - p_{k-1})^{n_{k+}}.$$

同样的调整适用于任何类型的分组频率数据. 假设频率在 3~5 间有 5 个观测, 其似然函数为 $(p_3 + p_4 + p_5)^5$.

例 12.53 利用表 12-9① 的数据确定 Poisson 分布的最大似然估计.

表 12-9 例 12.53 的数据

索赔数 (天)	观测数 (天)
0	47
1	97
2	109
3	62
4	25
5	16
6+	9

① 除了将 6 次以上的数据合并之外, 这里的数据与例 13.15 一样.

解 似然函数为

$$L = p_0^{47} p_1^{97} p_2^{109} p_3^{62} p_4^{25} p_5^{16} (1 - p_0 - p_1 - p_2 - p_3 - p_4 - p_5)^9,$$

若表示为 λ 的函数则有些复杂. 而且令导数为零进行求解需要数值方法. 此时可考虑直接用数值方法求似然函数的最大值. 首先可以假设所有 9 个观测都恰好为 6, 然后用样本均值作为一个合理的初值. 当然, 这将低估真正的最大似然概率, 但应该是一个好的出发点. 对于这个特定的例子, 最大似然估计是 $\hat{\lambda} = 2.022\,6$, 非常接近基于所有索赔次数记录所得到的估计. □

12.5.2 负二项分布

矩估计方程为

$$r\beta = \frac{\sum_{k=0}^{\infty} k n_k}{n} = \bar{x}, \tag{12.14}$$

$$r\beta(1+\beta) = \frac{\sum_{k=0}^{\infty} k^2 n_k}{n} - \left(\frac{\sum_{k=0}^{\infty} k n_k}{n}\right)^2 = s^2, \tag{12.15}$$

解为 $\hat{\beta} = (s^2/\bar{x}) - 1$ 和 $\hat{r} = \bar{x}/\hat{\beta}$. 注意这个方差估计是除以 n 得到的, 而不是 $n-1$, 这样的计算在矩方法估计中很常见, 但不是必要的. 同样注意, 如果 $s^2 < \bar{x}$, β 的估计将为负值, 这是不可以接受的值.

例 12.54 (续例 12.52) *使用矩方法估计负二项分布的参数.*

解 样本均值和样本方差分别为 2.5 和 3.05(请读者自己验证), 参数估计为 $\hat{r} = 11.364$ 和 $\hat{\beta} = 0.22$. □

与具有相同均值的 Poisson 分布相比较, 可以发现 β 是对 "额外的 Poisson" 变动的度量. $\beta = 0$ 意味着没有额外的 Poisson 变动, 而 $\beta = 0.22$ 时可以推出与具有相同均值的 Poisson 分布比较, 方差增加了 22%.

现在考虑最大似然估计. 负二项分布的对数似然函数为两个参数 β 和 r 的函数

$$\begin{aligned} l &= \sum_{k=0}^{\infty} n_k \ln p_k \\ &= \sum_{k=0}^{\infty} n_k \left[\ln \binom{r+k-1}{k} - r\ln(1+\beta) + k\ln\beta - k\ln(1+\beta)\right]. \end{aligned}$$

为了求对数似然函数的最大值, 分别对每个参数求导并令导数为零, 然后解参数. 对数似然函数的导数为

$$\frac{\partial l}{\partial \beta} = \sum_{k=0}^{\infty} n_k \left(\frac{k}{\beta} - \frac{r+k}{1+\beta}\right), \tag{12.16}$$

$$
\begin{aligned}
\frac{\partial l}{\partial r} &= -\sum_{k=0}^{\infty} n_k \ln(1+\beta) + \sum_{k=0}^{\infty} n_k \frac{\partial}{\partial r} \ln \frac{(r+k-1)\cdots r}{k!} \\
&= -n\ln(1+\beta) + \sum_{k=0}^{\infty} n_k \frac{\partial}{\partial r} \ln \prod_{m=0}^{k-1}(r+m) \\
&= -n\ln(1+\beta) + \sum_{k=0}^{\infty} n_k \frac{\partial}{\partial r} \sum_{m=0}^{k-1} \ln(r+m) \\
&= -n\ln(1+\beta) + \sum_{k=1}^{\infty} n_k \sum_{m=0}^{k-1} \frac{1}{r+m}. \qquad (12.17)
\end{aligned}
$$

令这些等式为零得到

$$\hat{\mu} = \hat{r}\hat{\beta} = \frac{\sum_{k=0}^{\infty} k n_k}{n} = \bar{x}, \qquad (12.18)$$

$$n\ln(1+\hat{\beta}) = \sum_{k=1}^{\infty} n_k \left(\sum_{m=0}^{k-1} \frac{1}{\hat{r}+m} \right). \qquad (12.19)$$

注意均值的最大似然估计为样本均值 (由定义, 与矩方法估计相同). 等式 (12.18) 和等式 (12.19) 可以用数值方法求解. 将 (12.19) 式中的 $\hat{\beta}$ 用 $\hat{\mu}/\hat{r}$ 替换得到

$$H(\hat{r}) = n\ln\left(1+\frac{\bar{x}}{\hat{r}}\right) - \sum_{k=1}^{\infty} n_k \left(\sum_{m=0}^{k-1} \frac{1}{\hat{r}+m} \right) = 0. \qquad (12.20)$$

如果 (12.15) 式的右端大于 (12.14) 式的右端, 可以证明 (12.20) 式存在唯一解. 否则, 负二项分布模型可能并不是一个好的模型, 因为样本方差没有超过样本均值.①

方程 (12.20) 可以用 Newton-Raphson 方法对 $\hat{r}$ 求数值解. 第 k 次迭代的方程为

$$r_k = r_{k-1} - \frac{H(r_{k-1})}{H'(r_{k-1})}.$$

常用的初值 r_0 为 r 的矩估计值. 当然, 任何数值求根方法 (如二分法, 切线法) 都可以使用.

这时的对数似然函数是一个二元函数. 它可以直接使用附录 F 中的方法求最大值. 对于完整数据的负二项分布情形, 因为已知均值的估计量为样本均值, 设 $\beta = \bar{x}/r$ 将使这个问题退化为一元情形.

例 12.55 对于例 12.52 中的数据确定负二项分布参数的最大似然估计.

解 似然函数的最大值点为 $\hat{r} = 10.965\,0$ 和 $\hat{\beta} = 0.227\,998$. □

① 也就是说, 当样本方差小于或者等于样本均值的时候, 对数似然函数不存在最大值. 随着 r 增加到无穷 β 减少到 0 这个函数会持续增加, 并保持乘积为常数. 实际上说明, 与数据匹配最好的模型是 Poisson 分布的极限情况 — 负二项分布.

例 12.56 Tröbliger[130] 通过调查一类机动车保单的 23 589 个机动车司机一年时间中出现事故的次数来研究他们的驾驶习惯, 表 12-10 分别给出由 Poisson 分布和负二项分布拟合的数据结果. 根据所给的信息, 确定哪个分布模型较优.

表 12-10 机动车索赔频率的两个模型

索赔数 (年)	驾驶员数	Poisson 期望	负二项期望
0	20 592	20 420.9	20 596.8
1	2 651	2 945.1	2 631.0
2	297	212.4	318.4
3	41	10.2	37.8
4	7	0.4	4.4
5	0	0.0	0.5
6	1	0.0	0.1
7+	0	0.0	0.0
参数		λ=0.144 220	r =1.117 90
			β =0.129 010
对数似然值		−10 297.84	−10 223.42

解 索赔数的期望值可以用样本容量 (23 589) 乘以模型对应的概率得到. 显然负二项分布的概率得到的期望值更加接近观测值. 另外, 负二项对数似然函数的最大值显著地大于 Poisson 分布的值. 第 13 章将讨论模型选择的正规过程 (包括显著大). 尽管如此, 这里负二项分布模型的优势是显而易见的. □

12.5.3 二项分布

二项分布有两个参数, m 和 q. 通常 m 的值是已知和确定的. 在这种情形下, 只有一个参数 q 是需要估计的. 在很多保险情景中, q 被解释为某种事件发生的概率, 如死亡或残疾, q 的一般估计为

$$\hat{q}=\frac{\text{事件发生数}}{\text{最大可能的事件数}},$$

当 m 已知时, $\hat{q}$ 即为矩方法估计量.

在本章前面例子中的频率数据情形中, 参数 m 为最大的可能观测量, 这个值可以是已知和确定的或者是未知的. 但在任何情形下, m 都不能小于最大观测量. 对数似然函数为

$$\begin{aligned} l&=\sum_{k=0}^{m} n_k \ln p_k \\ &=\sum_{k=0}^{m} n_k \left[\ln\binom{m}{k}+k\ln q+(m-k)\ln(1-q)\right]. \end{aligned}$$

当 m 是已知和确定的时, 只需要对 q 最大化 l.

$$\frac{\partial l}{\partial q}=\frac{1}{q}\sum_{k=0}^{m}kn_k-\frac{1}{1-q}\sum_{k=0}^{m}(m-k)n_k.$$

令其等于零, 得到

$$\hat{q}=\frac{1}{m}\frac{\sum_{k=0}^{m}kn_k}{\sum_{k=0}^{m}n_k},$$

即为已观测事件的比例. 对于矩方法, 当 m 确定时, q 的估计量和最大似然估计一样, 因为矩方程是

$$mq=\frac{\sum_{k=0}^{m}kn_k}{\sum_{k=0}^{m}n_k},$$

当 m 未知时, q 的最大似然估计量为

$$\hat{q}=\frac{1}{\hat{m}}\frac{\sum_{k=0}^{\infty}kn_k}{\sum_{k=0}^{\infty}n_k},\tag{12.21}$$

其中 $\hat{m}$ 为 m 的最大似然估计. 关于 m 和 q 的最大似然估计的一个简单方法是对不同的 m 值给出似然函数记录如下.

第 1 步: 首先令 $\hat{m}$ 等于最大观测值.

第 2 步: 由 (12.21) 式得到 $\hat{q}$.

第 3 步: 计算相应的对数似然函数.

第 4 步: 将 $\hat{m}$ 增加 1.

第 5 步: 重复步骤 2~4 直到取得最大值.

对于负二项分布, 并不一定存在一对参数使似然函数最大化. 特别地, 如果样本均值小于或者等于样本方差, 上面的过程随着 $\hat{m}$ 的增加会使对数似然函数持续增加. 将趋于一个 Poisson 模型, 这一点可以由例 12.52 的数据验证.

例 12.57 表 12-11 给出了 15 160 个保单每年的索赔次数. 求矩方法和最大似然估计量.

解 样本的均值和方差分别为 0.985 422 和 0.890 355. 方差小于均值, 说明二项分布是一个合理的分布模型. 用矩方法得到

$$mq=0.985\ 422,$$

$$mq(1-q)=0.890\ 355.$$

因此, $\hat{q}=0.096\ 474$ 和 $\hat{m}=10.214\ 40$. 但是, m 只能取整数值, 四舍五入后 $m=10$. 然后由第一个矩方程调整估计值为 $\hat{q}=0.098\ 542\ 2$. 这么做会使模型的方差和样本的方差不同, 因为 $10(0.098\ 542\ 2)(1-0.098\ 542\ 2)=0.888\ 316$. 由此看出使用矩方法估计整数值参数的缺陷.

表 12-11 每个保单的索赔次数

每个保单的索赔数	保单数
0	5 367
1	5 893
2	2 870
3	842
4	163
5	23
6	1
7	1
8+	0

现在我们尝试最大化似然函数. 根据数据 $m \geqslant 7$, 如果 m 已知只有 q 需要估计. 如果 m 未知可以从 7 开始增加 m 的值, 对每个确定的 m 值得到最大似然估计, 直至达到最大的似然概率. 这些结果列在表 12-12 中.

表 12-12 二项分布的似然函数值

$\hat{m}$	$\hat{q}$	- 对数似然值
7	0.140 775	19 273.56
8	0.123 178	19 265.37
9	0.109 491	19 262.02
10	0.098 542	19 260.98
11	0.089 584	19 261.11
12	0.082 119	19 261.84

最大的对数似然函数值出现在 $m = 10$. 如果 m 的值是事先未知的, 则参数的最大似然估计为 $\hat{m} = 10$ 和 $\hat{q} = 0.098\ 542\ 2$. 这与矩方法调整后的估计量一样. 这样的结果并不是对所有的数据集必然的. □

12.5.4 $(a, b, 1)$ 分布族

$(a, b, 1)$ 分布族的参数估计与 $(a, b, 0)$ 的原理相同.

假设数据和前面的例子具有相同的形式, 由 (4.13) 式, 似然函数为

$$L = (p_0^M)^{n_0} \prod_{k=1}^{\infty} (p_k^M)^{n_k} = (p_0^M)^{n_0} \prod_{k=1}^{\infty} [(1 - p_0^M) p_k^T]^{n_k}.$$

对数似然函数为

$$l = n_0 \ln p_0^M + \sum_{k=1}^{\infty} n_k [\ln(1 - p_0^M) + \ln p_k^T]$$

$$= n_0 \ln p_0^M + \sum_{k=1}^{\infty} n_k \ln(1 - p_0^M) + \sum_{k=1}^{\infty} n_k [\ln p_k - \ln(1 - p_0)],$$

其中最后一行由 $p_k^T = p_k/(1-p_0)$ 得到. $(a,b,1)$ 分布族的三个参数为 p_0^M, a 和 b, 这里的 a 和 b 将决定 $p_1, p_2, \cdots$.

可见

$$l = l_0 + l_1,$$

其中

$$l_0 = n_0 \ln p_0^M + \sum_{k=1}^{\infty} n_k \ln(1 - p_0^M),$$

$$l_1 = \sum_{k=1}^{\infty} n_k [\ln p_k - \ln(1 - p_0)],$$

这里 l_0 只依赖参数 p_0^M, 而 l_1 独立于参数 p_0^M 只依赖 a 和 b. 这使最大化过程变得简单, 因为

$$\frac{\partial l}{\partial p_0^M} = \frac{\partial l_0}{\partial p_0^M} = \frac{n_0}{p_0^M} - \sum_{k=1}^{\infty} \frac{n_k}{1 - p_0^M} = \frac{n_0}{p_0^M} - \frac{n - n_0}{1 - p_0^M},$$

得到

$$\hat{p}_0^M = \frac{n_0}{n},$$

它为在零点的观测比例. 这是一个自然的估计量, 因为 p_0^M 表示观测值取零的概率.

同样地, 应用似然函数分解提供的便利, 参数 a 和 b 的估计是独立于 p_0^M 的. 注意, 尽管 a 和 b 为参数, 最大化不应该完全基于它们. 因为并非 a 和 b 的所有值都会产生可接受的概率分布 [①]. 对于零点调整的 Poisson 分布, 对数似然函数的相关部分为

$$\begin{aligned} l_1 &= \sum_{k=1}^{\infty} n_k \left[\ln \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!} - \ln(1 - \mathrm{e}^{-\lambda}) \right] \\ &= -(n - n_0)\lambda + \left(\sum_{k=1}^{\infty} k n_k \right) \ln \lambda - (n - n_0) \ln(1 - \mathrm{e}^{-\lambda}) + c \\ &= -(n - n_0)[\lambda + \ln(1 - \mathrm{e}^{-\lambda})] + n\bar{x} \ln \lambda + c, \end{aligned}$$

这里 $\bar{x} = \frac{1}{n}\sum_{k=0}^{\infty} k n_k$ 是样本均值, $n = \sum_{k=0}^{\infty} n_k$, c 独立于 λ. 因此

$$\frac{\partial l_1}{\partial \lambda} = -(n - n_0) - (n - n_0)\frac{\mathrm{e}^{-\lambda}}{1 - \mathrm{e}^{-\lambda}} + n\frac{\bar{x}}{\lambda}$$

---

① 因为最大似然估计在参数变化下是不变的, 所以任何参数化的表示都可以求得最大值. 但是, 在有界区域求最大值时变得有些困难, 因为数值方法很难限制范围, 而分析方法会因为不可微而失效. 因此, 通常是针对特殊的分布族成员来完成估计, 比如 Poisson 分布.

$$= -\frac{n - n_0}{1 - e^{-\lambda}} + \frac{n\bar{x}}{\lambda}.$$

令其等于零, 得到

$$\bar{x}(1 - e^{-\lambda}) = \frac{n - n_0}{n}\lambda. \tag{12.22}$$

将等式两边分别看成 λ 的函数, 通过绘图可知, 如果 $n_0 > 0$, 恰好存在两个根：一个是 $\lambda = 0$; 另外一个是 $\lambda > 0$. 方程 (12.22) 可以用数值方法求解得到 $\hat{\lambda}$. 注意, 因为 $\hat{p}_0^M = n_0/n$ 和 $p_0 = e^{-\lambda}$, (12.22) 式可以改写为

$$\bar{x} = \frac{1 - \hat{p}_0^M}{1 - p_0}\lambda. \tag{12.23}$$

因为 (12.23) 式的右端是零点调整的 Poisson 分布的理论均值 (用 p_0^M 代替 $\hat{p}_0^M$), (12.23) 式是一个矩方程. 由此, 另一种方法将得到与最大似然估计相同的结果：令 p_0^M 等于样本在零点处的比例并且理论均值等于样本均值. 这说明, 若取零点概率为零观测的比例并利用低阶矩等式, 可以通过修正矩方法得到似然函数最大值数值求解的初值. 因为最大似然估计方法有比较好的渐近性质, 更可取的方法是只用修正矩方法作为初值.

为了得到最大似然估计量 λ 的渐近方差估计值, 容易得到

$$\frac{\partial^2 l_1}{\partial \lambda^2} = (n - n_0)\frac{e^{-\lambda}}{(1 - e^{-\lambda})^2} - \frac{n\bar{x}}{\lambda^2},$$

和期望值 $E(\bar{x}) = (1 - p_0^M)\lambda/(1 - e^{-\lambda})$. 最后, p_0^M 可以用它的估计量 n_0/n 来代替. 由于分子 n_0 服从二项分布得到 $\hat{p}_0^M$ 的方差为 $p_0^M(1 - p_0^M)/n$.

对于零点调整的二项分布, 有

$$\begin{aligned}
l_1 &= \sum_{k=1}^{m} n_k \left\{ \ln \left[\binom{m}{k} q^k (1 - q)^{m-k} \right] - \ln[1 - (1 - q)^m] \right\} \\
&= \left(\sum_{k=1}^{m} k n_k \right) \ln q + \sum_{k=1}^{m} (m - k) n_k \ln(1 - q) \\
&\quad - \sum_{k=1}^{m} n_k \ln[1 - (1 - q)^m] + c \\
&= n\bar{x} \ln q + m(n - n_0) \ln(1 - q) - n\bar{x} \ln(1 - q) \\
&\quad -(n - n_0) \ln[1 - (1 - q)^m] + c.
\end{aligned}$$

这里 c 不依赖于 q, 并且

$$\frac{\partial l_1}{\partial q} = \frac{n\bar{x}}{q} - \frac{m(n - n_0)}{1 - q} + \frac{n\bar{x}}{1 - q} - \frac{(n - n_0)m(1 - q)^{m-1}}{1 - (1 - q)^m}.$$

令上式为零得到

$$\overline{x}=\frac{1-\hat{p}_0^M}{1-p_0}mq, \tag{12.24}$$

其中 $p_0=(1-q)^m$. 这个方程使得理论均值和样本均值匹配.

如果 m 是已知和确定的, p_0^M 的最大似然估计量仍然是

$$\hat{p}_0^M=\frac{n_0}{n}.$$

但是, 即便 m 是已知的, (12.24) 式必须使用数值方法求 q. 当 m 是未知并且是待估参数时, 上面的过程可以对 m 取不同的值进行, 直至得到似然函数的最大值.

零点调整的负二项 (或扩展的截断负二项) 分布有些复杂, 因为需要估计 3 个参数. 当然, p_0^M 的最大似然估计与前面相同仍然为 $\hat{p}_0^M=n_0/n$, 这样问题简化为只需估计 r 和 β. 与 r 和 β 有关的对数似然函数为

$$l_1=\sum_{k=1}^{\infty}n_k\ln p_k-(n-n_0)\ln(1-p_0). \tag{12.25}$$

因此有

$$\begin{aligned}l_1=&\sum_{k=1}^{\infty}n_k\ln\left[\binom{k+r-1}{k}\left(\frac{1}{1+\beta}\right)^r\left(\frac{\beta}{1+\beta}\right)^k\right]\\&-(n-n_0)\ln\left[1-\left(\frac{1}{1+\beta}\right)^r\right].\end{aligned} \tag{12.26}$$

要在 (r,β) 平面上最大化这个函数, 从而得到最大似然估计. 可以利用附录 F 中描述的数值方法求得. 初值可以通过修正矩方法得到, 令 $\hat{p}_0^M=n_0/n$, 并且令分布的前两阶矩与样本的前两阶矩相等. 一般来说这时使用原点矩 (原点的矩) 比中心矩更简单. 在实践中, 最大化 (12.25) 式比 (12.26) 式更方便, 因为可以利用递归的优势

$$p_k=p_{k-1}\left(a+\frac{b}{k}\right)$$

来计算 (12.25) 式. 这使得计算机程序简单一些.

零点截断的分布不需要估计零点的概率值 (因为已知它为 0), 其余参数仍然用零点调整后分布导出的相同公式进行估计.

例 12.58 表 12-13 为 Beard et al.[12] 中的数据, 试给出一个能够充分描述该数据的模型.

解 套用 Poisson 分布模型结果非常差, 发生一次事故的概率太高, 而较高事故数的概率又太低. 几何分布是可选的另一个单参数模型, 对数似然函数为

$$l=-n\ln(1+\beta)+\sum_{k=1}^{\infty}n_k\ln\left(\frac{\beta}{1+\beta}\right)^k$$

$$
\begin{aligned}
&= -n\ln(1+\beta) + \sum_{k=1}^{\infty} kn_k[\ln\beta - \ln(1+\beta)] \\
&= -n\ln(1+\beta) + n\bar{x}[\ln\beta - \ln(1+\beta)] \\
&= -(n+n\bar{x})\ln(1+\beta) + n\bar{x}\ln\beta,
\end{aligned}
$$

其中 $\bar{x} = \sum\limits_{k=1}^{\infty} kn_k/n$ 和 $n = \sum\limits_{k=0}^{\infty} n_k$.

表 12-13 Beard 数据的拟合分布

事故数	观测数	Poisson	几何	ZM Poisson	ZM 几何
0	370 412	369 246.9	372 206.5	370 412.0	370 412.0
1	46 545	48 643.6	43 325.8	46 432.1	46 555.2
2	3 935	3 204.1	5 043.2	4 138.6	3 913.6
3	317	140.7	587.0	245.9	329.0
4	28	4.6	68.3	11.0	27.7
5	3	0.1	8.0	0.4	2.3
6+	0	0.0	1.0	0.0	0.2
参数		λ :0.13 174	β : 0131 74	p_0^M : 0.879 34	p_0^M: 0.879 34
				λ :0.178 27	β :0.091 780
对数似然函数		−171 373	−171 479	−171 160	−171 133

求导得到最大值点

$$\hat{\beta} = \bar{x}.$$

由表可以直接看出零点调整的几何分布匹配数据优于其他 3 个模型. 例 13.16 还将进行正式分析. □

12.5.5 复合模型

对复合模型采用矩方法时, 前几阶理论矩与样本矩进行匹配, 然后通过解方程组可以得到矩估计量. 需要注意的是复合模型的参数个数是主分布和次分布的参数个数之和. 复合分布前两阶矩的理论值为

$$
\begin{aligned}
\mathrm{E}(S) &= \mathrm{E}(N)\mathrm{E}(M), \\
\mathrm{Var}(S) &= \mathrm{E}(N)\mathrm{Var}(M) + \mathrm{E}(M)^2\mathrm{Var}(N).
\end{aligned}
$$

这些结果已在第 6 章中得到. 复合 Poisson 分布的前三阶矩由 (4.27) 式给出.

复合模型的最大似然估计仍然按以前的方法进行, 最大化的对数似然函数为

$$l = \sum_{k=0}^{\infty} n_k \ln g_k.$$

当 g_k 为复合分布的概率时, 仍然可以用数值方法求最大值. 这个对数似然函数的一阶导数、二阶导数可以通过对和式中的对数似然函数的最大值点的近似微分得到.

例 12.59 *确定 Poisson 零截断几何分布的性质. 也称该分布为 Polya-Aeppli 分布.*

解 对于零截断几何分布, 概率生成函数为

$$P_2(z)=\frac{[1-\beta(z-1)]^{-1}-(1+\beta)^{-1}}{1-(1+\beta)^{-1}}.$$

因此 Polya-Aeppli 的概率生成函数为

$$\begin{aligned}P(z)=P_1[P_2(z)]&=\exp\left(\lambda\left\{\frac{[1-\beta(z-1)]^{-1}-(1+\beta)^{-1}}{1-(1+\beta)^{-1}}-1\right\}\right)\\&=\exp\left\{\lambda\frac{[1-\beta(z-1)]^{-1}-1}{1-(1+\beta)^{-1}}\right\}.\end{aligned}$$

均值为

$$P'(1)=\lambda(1+\beta).$$

方差为

$$P''(1)+P'(1)-[P'(1)]^2=\lambda(1+\beta)(1+2\beta).$$

这意味着: $\mathrm{E}(N)=\mathrm{Var}(N)=\lambda, \mathrm{E}(M)=1+\beta$, 且 $\mathrm{Var}(M)=\beta(1+\beta)$, 则

$$\begin{aligned}\mathrm{E}(S)&=\lambda(1+\beta),\\\mathrm{Var}(S)&=\lambda\beta(1+\beta)+\lambda(1+\beta)^2=\lambda(1+\beta)(1+2\beta).\end{aligned}$$

由定理 4.51, 在零点的概率为

$$g_0=P_1(0)=\mathrm{e}^{-\lambda}.$$

接下来的 g_k 值可以通过复合 Poisson 递归简单计算

$$g_k=\frac{\lambda}{k}\sum_{j=1}^{k}jf_jg_{k-j},\quad k=1,2,3,\cdots,\tag{12.27}$$

其中 $f_j=\beta^{j-1}/(1+\beta)^j, j=1,2,\cdots$. 对于任意给定的 λ 和 β, 可以非常简单的计算出对数似然值. □

对于例 13.17 提供的数据集, Polya-Aeppli 分布是一个较好的选择.

另一个有用的复合 Poisson 分布是 Poisson 扩展截断负二项 (Poisson-ETNB) 分布. 尽管不在意次分布为调整的还是截断的, 我们还是更喜欢截断型, 因为这时的

参数 r 可以扩展①. 一些特例: $r=1$ 为 Poisson 几何 (也称做 Polya-Aeppli); $r\to 0$ 为 Poisson 对数 (负二项); $r=-0.5$ 为 Poisson 逆高斯. 最后一个的名字与其他的不一致, 因为逆高斯分布是一个混合分布 (见 4.6.9 节). Poisson 逆高斯分布将较好的拟合例 13.18 提供的数据集.

12.5.6 最大似然估计风险暴露水平的作用

在 4.6.11 节讨论了离散分布中风险暴露水平的作用. 当数据为团体数据时, 最大似然估计仍然是可能的. 下面以 Poisson 分布为例说明这一点.

例 12.60 *根据表 12-14 中的数据确定 Poisson 参数的最大似然估计.*

表 12-14 机动车年索赔数

年	暴露数	索赔数
1986	2 145	207
1987	2 452	227
1988	3 112	341
1989	3 458	335
1990	3 698	362
1991	3 872	359

解 令 λ 为每个风险暴露个体的 Poisson 参数. 如果第 k 年有 e_k 个风险暴露, 则索赔次数服从参数为 λe_k 的 Poisson 分布. 若 n_k 表示第 k 年的索赔次数, 似然函数为

$$L=\prod_{k=1}^{6}\frac{\mathrm{e}^{-\lambda e_k}(\lambda e_k)^{n_k}}{n_k!}.$$

最大似然估计为

$$l=\ln L=\sum_{k=1}^{6}[-\lambda e_k+n_k\ln(\lambda e_k)-\ln(n_k!)],$$

$$\frac{\partial l}{\partial\lambda}=\sum_{k=1}^{6}(-e_k+n_k\lambda^{-1})=0,$$

$$\hat{\lambda}=\frac{\sum_{k=1}^{6}n_k}{\sum_{k=1}^{6}e_k}=\frac{1\ 831}{18\ 737}=0.097\ 72.$$

□

在这个例子中, 答案与我们的预期相同, 为每个风险暴露的平均索赔数目. 这种方法对 $(a,b,0)$②和所有复合分布类都适用. 但必须谨慎地解释模型. 例如, 在采

① 这一点并不与定理 4.54 矛盾. 当 $-1<r<0$ 时, 改变零点的概率不会产生新的分布. 在零点没有概率将使分布变为平凡的 $(a,b,0)$ 负二项次分布.

② 对于二项分布, 仍然存在 m 必须为整数这个问题.

用负二项分布时，假设每个风险单位由负二项分布产生索赔，这与假设总索赔数服从负二项分布不同，后者的每个个体服从参数不同的 Poisson 分布.

习题

12.94 假设二项分布参数 m 已知. 考虑 q 的最大似然估计量.

(a) 证明这个最大似然估计量是无偏的.

(b) 给出最大似然估计的方差.

(c) 证明定理 12.13 给出的渐近方差与 (b) 的结果相同.

(d) 使用 9.3 节的 (9.4) 式给出一个简单的置信区间公式，方差项用 $\hat{q}$ 替换 q.

(e) 采用和 11.2 节例 11.12 相同的方法，由 (9.3) 式得到一个更复杂的置信区间公式.

12.95 利用 (12.18) 式给出几何分布 β 的最大似然估计量. 并给出最大似然估计量的方差，同时验证它与定理 12.13 中的渐近方差相同.

12.96 现有 10 000 个风险暴露的保单组，索赔数如表 12-15 所示.

表 12-15 习题 12.96 的数据

索赔数	保单数
0	9 048
1	905
2	45
3	2
4+	0

(a) 采用 Poisson 分布确定 λ 的最大似然估计和 95% 置信区间.

(b) 采用几何分布确定 β 的最大似然估计和 95% 置信区间.

(c) 采用负二项分布确定参数 r 和 β 的最大似然估计.

(d) 假设 $m = 4$，确定二项分布参数 q 的最大似然估计.

(e) 用习题 12.94 中 (d) 和 (e) 的方法构造 q 的 95% 置信区间.

(f) 通过计算似然函数值，确定 m 和 q 的最大似然估计.

12.97 某机动车辆保单将为其任何驾驶者 (投保的和未投保的) 的交通事故提供保险赔付. 1 000 个保单的数据如表 12-16 所示.

表 12-16 习题 12.97 的数据

索赔数	已投保的驾驶员数	未投保的驾驶员数
0	901	947
1	92	50
2	5	2
3	1	1
4	1	0
5+	0	0

(a) 采用 Poisson 模型，分别对变量 N_1 = 已投保驾驶者的索赔数和 N_2 = 未投保驾驶

者的索赔数确定 λ 的最大似然估计.

(b) 假设 N_1 和 N_2 独立. 使用定理 4.37 确定 $N = N_1 + N_2$ 的模型.

12.98 另一种计算习题 12.97 中 N 的模型的方法为, 只记录 1 000 个保单中的索赔总数而不区分是否为已投保驾驶员造成的事故. 如表 12-17 所示.

表 12-17 习题 12.98 的数据

索赔数	保单数
0	861
1	121
2	13
3	3
4	1
5	0
6	1
7+	0

(a) 采用 Poisson 模型确定 λ 的最大似然估计.

(b) (a) 部分的答案与前一题 (c) 部分的答案相同, 并证明这个结论是一般性的.

(c) 对于几何分布确定 β 的最大似然估计.

(d) 对于负二项分布确定 r 和 β 的最大似然估计.

(e) 假设 $m = 7$. 确定二项分布参数 q 的最大似然估计.

(f) 通过计算似然函数值, 确定 m 和 q 的最大似然估计.

12.99 表 12-18 中的数据为某中老年团体保险计划在一年内发生的处方数.

(a) 对于 Poisson 模型确定 λ 的最大似然估计.

(b) 对于几何分布确定 β 的最大似然估计, 然后确定 β 的 95% 置信区间.

(c) 对于负二项模型确定 r 和 β 最大似然估计.

表 12-18 习题 12.99 的数据

处方数	发生次数	处方数	发生次数
0	82	16～20	40
1～3	49	21～25	38
4～6	47	26～35	52
7～10	47	36～	91
11～15	57		

12.6 二元模型

12.6.1 引言

有时考虑相互依赖的二元变量会是更适合的模型. 例如联合生命年金或者寿

险的情形, 保险赔付的时间依赖于第一个或者第二个个体的死亡. 因为这些个体通常是相关的 (典型的如夫妇), 所以死亡时间相互依赖. 另一个例子, 为巨灾险中与损失赔付直接相关的费用数据, 称为已分摊的费用调整损失 (ALAE), 显然, 损失本身的赔付和 ALAE 通常有很强的正相关性.

有许多关于二元和多变量模型的文献, 包括 Hutchinson and Lai([64]), Kotz, Balakrishnan,and Johnson([80]) 和 Mardia([89]). 但是, 大多数模型的边缘分布并不是精算实务感兴趣的或者其参数不适合精算模型的特性 (例如, 一个二元 gamma 分布要求 X 和 Y 具有相同的 α 值). 一个例外是二元对数正态分布, 其对数服从二元正态分布.

由已知的边缘分布构造二元模型是我们更感兴趣和更具有实际价值的问题. 例如, 已知损失服从 Pareto 分布, ALAE 服从 gamma 分布, 则边缘数据模型的参数都是可以估计的 (模型是确定的). 接着通过引入两个变量之间的相关性, 可以产生一个二元分布. 在可选的方法中, 耦合方法得到了精算文献的广泛关注, 也是本节唯一介绍的方法.

12.6.2 耦合函数

耦合分布是采用一个耦合函数进行构造的, 这个函数本身必须为单位正方形上的一个正规的二元分布函数, 其边缘分布为均匀分布. 用 $F_X(x)$ 和 $F_Y(y)$ 表示两个边缘分布函数, $C(u,v)$ 表示耦合函数. 由这三个函数构造的二元分布为

$$F_{X,Y}(x,y)=C[F_X(x),F_Y(y)].$$

一个简单的但基本上无用的耦合函数是 $C(u,v)=uv$, 由此构造的二元分布函数为 $F_{X,Y}(x,y)=F_X(x)F_Y(y)$, 相互独立变量的联合分布就是这个表达式.

文献 [42] 对此进行了较好的一般性介绍, 与精算相关的介绍可以参见 [40]. Fees, Carriere and Valdez[39] 利用 Frank 耦合研究了联合生存时间问题. 文献 [78] 讨论了下面例子的扩展情形. 最后两篇文献说明在不同的删失和截断数据情形下如何构造似然函数.

例 12.61 *表 12-19 记录了 24 个索赔和相应的 ALAE. 设边缘分布为 Pareto 分布, 使用 Frank 耦合确定联合分布的模型.*

解 Frank 耦合函数为 (其中 $\log_\alpha$ 表示以 α 为底的对数)

$$C(u,v)=\log_\alpha\left[1+\frac{(\alpha^u-1)(\alpha^v-1)}{\alpha-1}\right], \tag{12.28}$$

这里的参数 α 用来控制两个变量之间的关联性. α 的值小于 1 表示正向关联性, 值大于 1 表示反向关联, 等于 1 表示独立. 如果令 β 和 θ 为 X 的 Pareto 边缘分布

的参数 (这里 θ 是尺度参数), 令 γ 和 τ 为 Y 的 Pareto 边缘分布的参数 (τ 是尺度参数), 二元分布函数为

表 12-19 24 个损失及其 ALAE 的记录

损失额	ALAE	损失额	ALAE
1 500	301	11 750	2 530
2 000	3 043	12 500	165
2 500	415	14 000	175
2 500	4 940	14 750	28 217
4 500	395	15 000	2 072
5 000	25	17 500	6 328
5 750	34 474	19 833	212
7 000	50	30 000	2 172
7 000	10 593	33 033	7 845
7 500	50	44 887	2 178
9 000	406	62 500	12 251
10 000	1 174	210 000	7 357

$$F(x,y)=\log_\alpha\left\{1+\frac{[\alpha^{1-(1+x/\theta)^{-\beta}}-1][\alpha^{1-(1+y/\tau)^{-\gamma}}-1]}{\alpha-1}\right\}.$$

对 x 和 y 求偏导得到联合密度函数

$$f(x,y)=\frac{\begin{matrix}(\alpha-1)\dfrac{\beta\gamma}{\theta\tau}\alpha^{2-(1+x/\theta)^{-\beta}-(1+y/\tau)^{-\gamma}}\\ \times(1+x/\theta)^{-\beta-1}(1+y/\tau)^{-\gamma-1}\ln\alpha\end{matrix}}{\{\alpha-1+[\alpha^{1-(1+x/\theta)^{\beta}}-1][\alpha^{1-(1+y/\tau)^{-\gamma}}-1]\}^2}.$$

4 个 Pareto 参数的初值可以通过两个边缘分布的最大似然估计得到. 考虑单纯形最大化方法得到估计量 $\hat{\alpha}=0.133\ 024$, $\hat{\beta}=2.598\ 89$, $\hat{\theta}=36\ 141.4$, $\hat{\gamma}=0.759\ 943$, $\hat{\tau}=803.839$. 显然是正向关联的并可以检验, 一种方法是使用第 13 章讨论的似然比检验, 但较小的样本量并不能充分地说明正向关联的程度. □

Genest[41] 讨论了 Frank 耦合的一些结果. 这里只介绍其中的两个结果. 在模拟 (X,Y) 数据对时, 可以先由边缘分布来模拟 X 的取值, 这可以通过标准的逆函数得到, 然后根据 Y 关于 $X=x$ 的条件概率分布来模拟 Y 的取值. 首先, 条件分布函数

$$F_{Y|X}(y|x)=\frac{(\partial/\partial x)F(x,y)}{f_X(x)}.$$

对于 Frank 耦合, 有

$$\frac{\partial}{\partial x}F(x,y)=f_X(x)\frac{\partial}{\partial u}C(u,v)|_{u=F_X(x),v=F_Y(y)}$$

$$=\frac{f_X(x)\alpha^{F_X(x)}[\alpha^{F_Y(y)}-1]}{\alpha-1+[\alpha^{F_X(x)}-1][\alpha^{F_Y(y)}-1]}.$$

为了模拟 Y 的条件值, 采用第 17 章讨论的求逆方法, 由一个 $(0,1)$ 上均匀分布的随机数 r 解方程

$$\frac{\alpha^{F_X(x)}[\alpha^{F_Y(y)}-1]}{\alpha-1+[\alpha^{F_X(x)}-1][\alpha^{F_Y(y)}-1]}=r.$$

对于 $F_Y(y)$ 得到

$$\alpha^{F_Y(y)}=1+\frac{r(\alpha-1)}{\alpha^{F_X(x)}(1-r)+r}$$

或

$$F_Y(y)=\frac{1}{\ln\alpha}\ln\left[1+\frac{r(\alpha-1)}{\alpha^{F_X(x)}(1-r)+r}\right].$$

右边是一个数值, 对 Y 的分布函数求逆解得模拟值.

还可以由下面的公式得到回归函数

$$\begin{aligned}\mathrm{E}(Y|X=x)&=\int[1-F_{Y|x}(y|x)]\mathrm{d}y\\&=\int\left\{1-\frac{\alpha^{F_X(x)}[\alpha^{F_Y(y)}-1]}{\alpha-1+[\alpha^{F_X(x)}-1][\alpha^{F_Y(y)}-1]}\right\}\mathrm{d}y,\end{aligned}$$

但很可能需要使用数值方法计算积分.

习题

12.100 考虑表 12-19 中的数据集, 使用 Frank 耦合的二元分布模型, 设边缘分布为逆指数分布.

12.7 协变量模型

12.7.1 引言

有时我们关心的随机变量可能会依赖于周围环境的特定属性. 例如, 生存时间可能与个体的年龄、性别、吸烟状况、血压、身高、体重等因素相关. 或者, 考虑机动车辆的年事故数, 这个随机变量的分布可能与机动车的驾驶里程、驾驶地区和主要驾驶员的状况如年龄、性别、婚姻状况和驾驶记录有关.

例 12.62 *假设认为机动车年事故数的分布与驾驶员的年龄、性别相关. 试给出 3 种建模方法.*

解 当然, 可选的模型有无限多个. 3 个可能选用的模型如下.

(1) 对于性别和年龄的每个组合分别建立模型. 对于每种组合分别收集数据和建模. 任何参数或者数据依赖模型都可以被选用.

(2) 构造一个完整的单参数模型. 作为一个例子, 可以假设事故次数服从参数为 λ 的 Poisson 分布. 可以认为 λ 依赖于年龄 x 和性别 (男性 $g=1$, 女性 $g=0$), 比如

$$\lambda = (\alpha_0 + \alpha_1 x + \alpha_2 x^2)\beta^g.$$

(3) 与数据依赖模型相似, 首先构造模型的密度、分布或者风险率函数. 然后用年龄和性别变量修正这些函数. 例如, 首先选择生存函数 $S_0(n)$, 则每个特定驾驶员的生存函数为

$$S(n|x,g) = [S_0(n)]^{(\alpha_0+\alpha_1 x+\alpha_2 x^2)\beta^g}. \qquad \square$$

尽管第一种方法没有什么问题, 但并不是很有意义, 它只是要求对于不同的组合一次次地重复建模过程. 第二个方法是一个单参数模型, 可以采用讨论过的各种技术分析, 显得比较简单. 第三个模型具有混合的特性, 需要更多的工作量实现.

当给定的个体不满足任何明显的分布模型时, 第三个模型将是不错的选择. 对于机动车的例子, Poisson 分布是一个合理的选择, 因此第二个模型可能是最好的方法. 如果变量是生存时间, 数据依赖的模型, 如生命表可能是恰当的选择.

第二个模型和第三个模型相对于第一个的优势是在某些组合的观测非常少的情形. 在这种情况下, 第二种模型和第三种模型的节约性原则使得有限的信息仍然很有价值. 例如, 要对性别/吸烟状况的 4 种组合建立从 20 岁到 99 岁的生命表. 使用模型 1 需要进行 320 个估计, 而使用模型 3 只需要 83 个估计.[①]

12.7.2 比例风险模型

Cox 比例风险模型是相对简单的方法.

定义 12.63 *给定某个基准风险率函数 $h_0(t)$ 以及某个体的相关变量值 $z_1,\cdots,z_p$, 则定义该个体的 Cox 比例风险模型的风险率函数如下*

$$h(x|\boldsymbol{z}) = h_0(x)c(\beta_1 z_1 + \cdots + \beta_p z_p) = h_0(x)c(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}),$$

其中 $c(y)$ 为任意取正值的函数, $\boldsymbol{z}=(z_1,\cdots,z_p)^{\mathrm{T}}$ 是 z 值 (称作协变量) 的列向量, $\boldsymbol{\beta}=(\beta_1,\cdots,\beta_p)^{\mathrm{T}}$ 为系数列向量.

这部分只使用一个函数 $c(y)=\mathrm{e}^y$. 这个函数的一个优点是它一定取正值. 模型的名字也是合适的, 因为当两个个体的风险率函数取定时, 两者的比例为常数. 也就是说, 一个个体的风险率函数与另一个的成比例. 剩下的就是估计基准风险率函数 $h_0(t)$ 和系数向量 $\boldsymbol{\beta}$.

---

① 对于 4 种组合中某个组合的生存函数需要 80 个估计. 其他的 3 种组合在已经得到的生存函数的指数上各增加了 1 个估计.

例 12.64 假设某房屋火灾保险的赔付以房屋价值的百分比表示, 依赖于房屋的使用年限和建筑材料 (木质或砖质). 对这个情形构造 Cox 比例风险率函数. 另说明相同年限房屋的木屋和砖屋的区别.

解 令 $z_1 =$ 使用年限 (非负整数), 如果建筑材料是木质 $z_2 = 1$, 如果建筑材料是砖质 $z_2 = 0$. 某个房屋的风险率函数为

$$h(x|z_1, z_2) = h_0(x)\mathrm{e}^{\beta_1 z_1 + \beta_2 z_2}.$$

这个模型的结果是, 如果不考虑房屋的使用年限, 则砖房与木房的风险是一样的. 对于两个使用年限为 z_1 的房屋, 有

$$h_{\text{木质}}(x) = h_0(x)\mathrm{e}^{\beta_1 z_1 + \beta_2} = h_{\text{砖质}}(x)\mathrm{e}^{\beta_2}.$$

对生存函数的作用为

$$\begin{aligned} S_{\text{木质}}(x) &= \exp\left[-\int_0^x h_{\text{木质}}(y)\mathrm{d}y\right] = \exp\left[-\int_0^x h_{\text{砖质}}(y)\mathrm{e}^{\beta_2}\mathrm{d}y\right] \\ &= [S_{\text{砖质}}(x)]^{\exp(\beta_2)}. \end{aligned}$$

□

可以通过参数模型也可以通过数据依赖模型来估计基准风险率函数. 然后就剩下参数了. 根据本书的基本原则, 我们将使用最大似然估计方法来估计 β_1 和 β_2. 下面列举了一个完全参数化的例子.

例 12.65 表 12-20 给出了火灾保险 10 次赔付的数据, 所有数值均为房屋价值的百分比. 用最大似然法估计 Cox 比例风险模型的参数, 分别使用指数分布和 beta 分布作为基准风险函数. 并已知这些保单没有免赔但有保单限额 (各保单不同).

表 12-20 火灾险的赔付数据

z_1	z_2	赔付
10	0	70
20	0	22
30	0	90*
40	0	81
50	0	8
10	1	51
20	1	95*
30	1	55
40	1	85*
50	1	93

* 赔付达到了保单限额

解 为了构造似然函数, 需要知道密度函数和生存函数. 令 $c_j = \exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z})$ 是第 j 个观测的 Cox 乘数. 则如前例所示: $S_j(x) = S_0(x)^{c_j}$, 其中 $S_0(x)$ 为基准分布. 密度函数为

$$
\begin{aligned}
f_j(x) &= -S_j'(x) = -c_j S_0(x)^{c_j-1} S_0'(x) \\
&= c_j S_0(x)^{c_j-1} f_0(x).
\end{aligned}
$$

对于指数分布, 有

$$
S_j(x) = [\mathrm{e}^{-x/\theta}]^{c_j} = \mathrm{e}^{-c_j x/\theta} \quad 且 \quad f_j(x) = \left(\frac{c_j}{\theta}\right) \mathrm{e}^{-c_j x/\theta}.
$$

对于 beta 分布, 有

$$
\begin{aligned}
S_j(x) &= [1-\beta(a,b;x)]^{c_j}, \\
f_j(x) &= c_j[1-\beta(a,b;x)]^{c_j-1} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1},
\end{aligned}
$$

其中 $\beta(a,b;x)$ 表示参数为 a 和 b 的贝塔分布的分布函数. [在 Excel® 中的函数名为 BETADIST(x,a,b)]. gamma 函数在 Excel® 中为 EXP(GAMMALN(a)). 对于没有达到保单限额的赔付, 其对似然函数的贡献为密度函数, 那些达到了限额的保单对似然函数的贡献为生存函数. 在两种情况下, 似然函数都变得相当复杂, 以至于不值得将其表示出来. 指数模型的参数估计为 $\hat{\beta}_1 = 0.003\ 19, \hat{\beta}_2 = -0.637\ 22$ 和 $\hat{\theta} = 0.740\ 41$, 似然函数的对数值为 $-6.137\ 9$. beta 模型估计值为 $\hat{\beta}_1 = -0.003\ 15, \hat{\beta}_2 = -0.778\ 47$, $\hat{a} = 1.037\ 06$ 和 $\hat{b} = 0.814\ 42$, 似然函数的对数值为 $-4.215\ 5$. 使用 Schwarz Bayesian 准则 (见 13.5.3 节), 需要调整第 4 个参数为 $\ln(10)/2 = 1.151\ 3$ 进行改进. beta 分布是较好的模型. 如果希望得到信息矩阵的估计, 唯一合理的策略是求对数似然函数的数值导数. □

另一个可选的方法是为基准风险率构造数据依赖模型. 令 $R(y_j)$ 表示观测值集, 其元素在未删失观测 y_j 的风险集中.[①] 这时不是计算真实的似然值, 而是考虑更易求得的偏似然值, 是一个条件值. 我们不问 "观测到 y_j 的概率是多少?" 而是问 "给定已经观测到一个不删失的数值 y_j, 每个保单具有这个观测值的概率是多少? 这是在等于或超过这个观测值的条件下进行的计算." 这个方法使我们可以独立于基准风险率来估计 β 系数. 这里的记号有些复杂, 令 j^* 表示产生没有删失的观测 y_j 的保单, 那么该保单对似然函数的贡献为

$$
\frac{f_{j^*}(y_j)/S_{j^*}(y_j)}{\sum_{i\in R(y_j)} f_i(y_j)/S_i(y_j)} = \frac{c_{j^*} f_0(y_j)/S_0(y_j)}{\sum_{i\in R(y_j)} c_i f_0(y_j)/S_0(y_j)} = \frac{c_{j*}}{\sum_{i\in R(y_j)} c_i}.
$$

① 回忆 11.1 节, $y_1, y_2, \cdots$ 表示未删失观测集合中有序的唯一值. 11.1 节还给出了风险集的定义.

例 12.66 对例 12.65 的数据使用偏似然法估计 β_1 和 β_2.

解 排序后的未删失数据为 $8, 22, 51, 55, 70, 81, 93$. 这些观测对似然函数的贡献列在表 12-21 中.

在 $\hat{\beta}_1 = -0.003\ 73$ 和 $\hat{\beta}_2 = -0.919\ 94$ 时, 乘积取得最大值, 其偏似然值的对数为 $-11.988\ 9$. 当 β_1 强制为 0 时, 最大值在 $\hat{\beta}_2 = -0.937\ 08$ 处, 偏似然值的对数为 $-11.996\ 8$. 在这个样本中没有证据表明, 房屋的使用年限对模型有任何影响. □

表 12-21 火灾险的似然数据

值	y	c	在 L 中占的比
8	8	$c_1 = \exp(50\beta_1)$	$\frac{c_1}{c_1+\cdots+c_{10}}$
22	22	$c_2 = \exp(20\beta_1)$	$\frac{c_2}{c_2+\cdots+c_{10}}$
51	51	$c_3 = \exp(10\beta_1 + \beta_2)$	$\frac{c_3}{c_3+\cdots+c_{10}}$
55	55	$c_4 = \exp(30\beta_1 + \beta_2)$	$\frac{c_4}{c_4+\cdots+c_{10}}$
70	70	$c_5 = \exp(10\beta_1)$	$\frac{c_5}{c_5+\cdots+c_{10}}$
81	81	$c_6 = \exp(40\beta_1)$	$\frac{c_6}{c_6+\cdots+c_{10}}$
85		$c_7 = \exp(40\beta_1 + \beta_2)$	
90		$c_8 = \exp(30\beta_1)$	
93	93	$c_9 = \exp(50\beta_1 + \beta_2)$	$\frac{c_9}{c_9+c_{10}}$
95		$c_{10} = \exp(20\beta_1 + \beta_2)$	

还有 3 个问题没有解决. 一个是估计基准风险率函数, 一个是处理存在多个观测有相同值的情形, 还有一个是估计量的方差的估计. 对于第二个问题, 文献中有些处理的方法. 先前提到的问题可以改述为 "已知有 s_j 个无删失的观测 y_j, 这 s_j 个保单具有这个观测值的概率是多少? 在等于或者超过这个观测值的条件下计算." 这个陈述的一个直接的解释是分子反映了得到 s_j 个观测值的概率. 分母根据 $R(y_j)$ 的所有 s_j 个元素的子集得到. 这需要大量的计算, Breslow 给出的一个简化版本是分别对待 s_j 中的每个观测, 但是对于分母它们都使用相同的风险集, 为了达到这个效果要求对上面介绍的算法不做任何改变.

例 12.67 在前面的例子中, 假设观测值为 81 的样本实际值为 70. 给出这两个观测对于偏似然函数的贡献.

解 使用例 12.66 的记号, 第一个观测 70 的贡献仍然是 $c_5/(c_5 + \cdots + c_{10})$. 但是, 第二个观测 70 的贡献为 $c_6/(c_5 + \cdots + c_{10})$. 注意分子没有变 (仍然是 c_6); 然而, 分母表明 $R(70)$ 中有 6 个观测. □

关于风险率函数的估计, 首先注意到累积风险率函数为

$$H(t|\boldsymbol{z}) = \int_0^t h(u|\boldsymbol{z})\mathrm{d}u = \int_0^t h_0(u)c\mathrm{d}u = H_0(t)c.$$

与 Nelson-Åalen 估计类似, 采用

$$\hat{H}_0(t) = \sum_{y_j \leqslant t} \frac{s_j}{\sum_{i \in R(y_j)} c_i}.$$

这里外面的求和是对所有小于或者等于 t 的未删失观测进行的. 分子为未删失值等于 y_j 的观测数目, 分母不再是对风险集的数求和, 而是对它们的 c 值求和. 和通常一样, 基准生存函数估计为 $\hat{S}_0(t) = \exp[-\hat{H}_0(t)]$.

例 12.68 继续前面的例子 (使用原始值), 估计基准生存函数, 然后再估计使用了 35 年的木屋索赔超过房子价值 80%的概率. 将得到的结果与前面由 beta 分布模型得到的值进行比较.

解 使用前面得到的估计, 10 个 c 值在表 12-22 中给出. 其中还包括了累积风险估计的跳跃值, 接着是累积风险函数本身的估计, 两个函数值均按照 y 的区间 (但不包括右端点) 进行计算.

表 12-22 火灾险的基准生存函数

值	y	c	跳的幅度	$\hat{H}_0(y)$	$\hat{S}_0(y)$
8	8	0.830 0	$\frac{1}{0.8300+\cdots+0.3699}$=0.159 7	0.159 7	0.852 4
22	22	0.928 2	$\frac{1}{0.9282+\cdots+0.3699}$=0.184 1	0.343 8	0.709 1
51	51	0.384 0	$\frac{1}{0.3840+\cdots+0.3699}$=0.222 0	0.565 8	0.567 9
55	55	0.356 4	$\frac{1}{0.3564+\cdots+0.3699}$=0.242 7	0.808 6	0.445 5
70	70	0.963 4	$\frac{1}{0.9634+\cdots+0.3699}$=0.265 7	1.074 3	0.341 5
81	81	0.861 5	$\frac{1}{0.8615+\cdots+0.3699}$=0.357 2	1.431 5	0.239 0
85		0.343 4			
90		0.894 2			
93	93	0.330 8	$\frac{1}{0.3308+0.3699}$=1.427 1	2.858 6	0.057 4
95		0.369 9			

对于题目描述的房屋情况, 有 $c = \exp[-0.003\,73(35) - 0.919\,94(1)] = 0.349\,77$, 估计概率为 $0.341\,5^{0.349\,77} = 0.686\,74$.

由 beta 分布可知, $\hat{S}_0(0.8) = 0.277\,32$ 和 $c = \exp[-0.003\,15(35) - 0.778\,47(1)] = 0.411\,18$, 得到的估计概率为 $0.277\,32^{0.411\,18} = 0.590\,15$. □

关于方差的估计, 偏似然函数的对数为

$$l(\boldsymbol{\beta}) = \sum_{j*} \ln \frac{c_{j*}}{\sum_{i \in R(y_j)} c_i},$$

其中是对所有无删失的观测求和. 对 β_g 求一阶偏导得到

$$\frac{\partial}{\partial \beta_g} l(\boldsymbol{\beta}) = \sum_{j^*} \left[\frac{1}{c_{j^*}} \frac{\partial c_{j^*}}{\partial \beta_g} - \frac{1}{\sum_{i \in R(y_j)} c_i} \frac{\partial}{\partial \beta_g} \sum_{i \in R(y_j)} c_i \right].$$

为了简化这个表达式, 注意到

$$\frac{\partial c_{j^*}}{\partial \beta_g}=\frac{\partial \mathrm{e}^{\beta_1 z_{j^*1}+\beta_2 z_{j^*2}+\cdots+\beta_p z_{j^*p}}}{\partial \beta_g}=z_{j^*g}c_{j^*},$$

这里 z_{j*g} 为给定 $j*$ 时的 z_g 值. 导数为

$$\frac{\partial}{\partial \beta_g}l(\boldsymbol{\beta})=\sum_{j^*}\left[z_{j*g}-\frac{\sum_{i\in R(y_j)}z_{ig}c_i}{\sum_{i\in R(y_i)}c_i}\right].$$

负二阶偏导为

$$\begin{aligned}&-\frac{\partial^2}{\partial\beta_h\beta_g}l(\boldsymbol{\beta})\\&=\sum_{j*}\left[\frac{\sum_{i\in R(y_j)}z_{ig}z_{ih}c_i}{\sum_{i\in R(y_j)}c_i}-\frac{\left(\sum_{i\in R(y_j)}z_{ig}c_i\right)\left(\sum_{i\in R(y_j)}z_{ih}c_i\right)}{\left(\sum_{i\in R(y_j)c_i}\right)^2}\right].\end{aligned}$$

将估计值代入后, 这些偏导数给出了信息矩阵的估计.

例 12.69 继续前面的例子, 求信息矩阵并估计协方差矩阵. 然后用此构造一个具有相同使用年限的木屋和砖屋的相对风险的 95%置信区间.

解 考虑对观测 $z_1=50$ 和 $z_2=1$ 的外面的求和计算. 风险集包括这个观测 (值为 93 和 $c=0.330\ 802$) 和 $z_1=20$, $z_2=1$ 的删失观测 (值为 95 和 $c=0.369\ 924$). 对 β_1 和 β_2 求导, 求和项为

$$\begin{aligned}&\frac{50(1)(0.330\ 802)+20(1)(0.369\ 924)}{0.330\ 802+0.369\ 924}\\&-\frac{[50(0.330\ 802)+20(0.369\ 924)][1(0.330\ 802)+1(0.369\ 924)]}{(0.330\ 802+0.369\ 924)^2}=0.\end{aligned}$$

对这些项目求和, 然后对于其他的偏导数作同样的处理, 得到信息阵和逆 - 协方差阵

$$I=\begin{bmatrix}1\ 171.054 & 5.976\ 519\\ 5.976\ 519 & 1.322\ 283\end{bmatrix},\quad \widehat{\mathrm{Var}}=\begin{bmatrix}0.000\ 874 & -0.003\ 95\\ -0.003\ 95 & 0.774\ 125\end{bmatrix}.$$

这两种情形的相对风险是比例 c 值. 对于一个使用年限为 x 的房屋, 木屋对砖屋的相对风险是 $\mathrm{e}^{z_1\beta_1+\beta_2}/\mathrm{e}^{z_1\beta_1}=\mathrm{e}^{\beta_2}$. β_2 的 95%置信区间为 $-0.919\ 94\pm1.96\sqrt{0.774\ 125}$ 或者 $(-2.644\ 4, 0.804\ 55)$. 对端点求指数则给出相对风险的置信区间 $(0.071\ 05, 2.235\ 7)$.

12.7.3 广义线性和加速失效模型

比例风险模型要求生存函数之间具有特殊的关系. 从精算的角度来说, 这可能不是最合理的, 因为很难解释风险率函数乘以一个常数的含义 (或者, 等价地, 对生

存函数求幂)[①]. 考虑将协变量与直接关心的量例如期望值建立联系, 可能更加有意义. 如标准多元回归模型的一般线性模型不一定合理, 因为它们通常趋向正态分布, 这并不适合大多数精算师所关心的现象. 广义线性模型放弃了正态的限制, 因此变得更实用. 文献 [90] 有较详细的讨论, 关于这个模型的精算论文包括 [47],[61],[93],[97]. 下面给出了这个模型的定义, 这个定义比通常的更具一般性.

定义 12.70 假设模型的参数为 μ 和 $\boldsymbol{\theta}$, μ 为均值、$\boldsymbol{\theta}$ 是其他参数组成的向量, 这个均值不依赖于其他的参数, 并且其他参数也不依赖均值. 设累积分布函数为 $F(x|\mu,\boldsymbol{\theta})$. 令 $\boldsymbol{z}$ 为某个体的协变量向量, $\boldsymbol{\beta}$ 为系数向量, $\eta(\mu)$ 和 $c(y)$ 为函数. 按照广义线性模型随机变量 X 的分布函数为

$$F(x|\boldsymbol{z},\boldsymbol{\theta})=F(x|\mu,\boldsymbol{\theta}),$$

其中 μ 满足 $\eta(\mu)=c(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z})$.

这个模型表明个体观测的均值通过特定的函数集合与协变量相关联. 正常情况下, 这些函数不含有参数, 而只是提供了一种更好的拟合或者保证只有合理的 μ 值被考虑.

例 12.71 证明一般的线性回归模型为广义线性模型的特例.

解 一般的线性回归, X 服从正态分布, $\mu=\mu$ 和 $\theta=\sigma^2$. η 和 c 都是恒等函数, 使得 $\mu=\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}$. □

这里介绍的模型比通常假设的 X 服从某些特殊分布的模型更具有一般性, 因为对于这些分布, 可以使用所有的回归分析工具, 如残差分析. 而含有广义线性模型的计算机程序也只限于这些分布.

我们讨论过的许多分布中, 有时均值并不是一个参数. 尽管也可以将其看作一个参数. 例如, 可以对 Pareto 分布参数化, 令 $\mu=\theta/(\alpha-1)$, 或者, 等价地, 将 θ 用 $\mu(\alpha-1)$ 代替, 则分布函数表示为

$$F(x|\mu,\alpha)=1-\left[\frac{\mu(\alpha-1)}{\mu(\alpha-1)+x}\right]^{\alpha},\quad \mu>0,\alpha>1.$$

注意在参数空间中对 α 的限制.

例 12.72 对于习题 12.65 的数据构造广义线性模型, 以 beta 分布作为损失模型.

解 beta 分布的参数见附录 A, 均值为 $\mu=a/(a+b)$, 另外一个参数为 $\theta=b$. 建立协变量与均值的关系的一个方法是令 $\eta(\mu)=\mu/(1-\mu)$ 和 $c(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z})=\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z})$. 解这些等式得到

$$\mu=\frac{\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z})}{1+\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z})}.$$

---

① 尽管如此, 在寿险中为了考虑已知的健康风险 (如肥胖) 在 q_x 上乘以一个常数的方法并不少见. 这与在风险率函数上乘一个常数没有多少不同.

求解前两个方程得到 $a = b\mu/(1-\mu) = b\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z})$. 对于未删失观测使用 $f(x)$ 作为似然因子, 对于删失观测使用 $1-F(x)$ 作为似然因子, 然后考虑最大似然估计. 对于每个观测, beta 分布直接使用参数 b 和由参数 b 以及该观测的协变量计算得到的参数 a. 因为没有基准分布, 表达式 $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}$ 一定包含常数项. 最大化似然函数得到的估计为 $\hat{b} = 0.577\ 5, \hat{\beta}_0 = 0.213\ 0, \hat{\beta}_1 = 0.001\ 8$ 和 $\hat{\beta}_2 = 1.094\ 0$. 与例 12.65 一样, 使用年限的影响可以忽略. 这个模型的一个优势是均值直接与协变量相关联. □

与广义线性模型具有相同思想的另一个模型是加速失效模型, 描述如下.

定义 12.73 *加速失效模型的定义如下*

$$S(x|\boldsymbol{z},\boldsymbol{\beta}) = S_0(x\mathrm{e}^{-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}}). \tag{12.29}$$

可以看出在给定均值时, 这就是广义线性模型, 首先注意到, (假设 $S(0)=1$),

$$\mathrm{E}(X|\boldsymbol{z},\boldsymbol{\beta}) = \int_0^\infty S_0(x\mathrm{e}^{-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}})\mathrm{d}x = \int_0^\infty \mathrm{e}^{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}} S_0(y)\mathrm{d}y = \exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z})E_0(X),$$

由此得知均值与协变量有关. 这个模型的名字的由来是因为协变量实际上改变了年龄. 一个年龄为 x, 协变量是 $\boldsymbol{z}$ 的人未来生存时间和一个年龄为 $x\mathrm{e}^{-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z}}$, $\boldsymbol{z}=0$ 的人的未来生存时间具有相同的分布. 如果基准分布有一个尺度参数, 那么协变量的作用为在尺度参数上乘一个常数. 因此, 如果 θ 是基准分布的一个尺度参数, 那么一个协变量为 $\boldsymbol{z}$ 的人与尺度参数为 $\exp(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{z})\theta$ 的人具有相同的生存分布. 与广义线性模型不同, 在使用这个模型之前并不要求均值存在.

例 12.74 *现有对 50 ~ 59 岁人群的死亡率研究, 数据包括性别和血压 100,125 或者 150. 对于这 6 种组合与 10 个年龄, 总计 1 000 个个体被观测, 死亡人数数据见表 12-23. 在 Gompertz 分布基础上估计加速失效模型的参数.*

解 Gompertz 的风险率函数为 $h(x) = Bc^x$, 生存函数为 $S_0(x) = \exp[-B(c^x-1)/\ln c]$ 作为基准函数. 令协变量 $z_1 = 0$ 表示男性, $z_1 = 1$ 表示女性, z_2 表示血压. 对于每个个体, 令 $\gamma = \exp(\beta_1 z_1 + \beta_2 z_2)$, 此时的加速失效模型为

$$S(x|\gamma) = S_0\left(\frac{x}{\gamma}\right) = \exp\left[-\frac{B(c^{x/\gamma}-1)}{\ln c}\right].$$

令 $c* = c^{1/\gamma}$, $B^* = B/\gamma$, 则

$$S(x|\gamma) = \exp\left[-\frac{B^*\gamma(c^{*x}-1)}{\gamma\ln c^*}\right] = \exp\left[-\frac{B^*(c^{*x}-1)}{\ln c^*}\right],$$

因此分布仍然为 Gompertz 分布, 新的参数如上所示. 对于每个年龄, 有

$$q_x|\gamma = 1 - \frac{S(x+1|\gamma)}{S(x|\gamma)} = 1 - \exp\left[-\frac{B^*c^{*x}(c^*-1)}{\ln c^*}\right].$$

表 12-23 例 12.74 的数据

年龄	男性 (0)			女性 (1)		
	100	125	150	100	125	150
50	13	12	85	3	12	49
51	11	21	95	7	13	53
52	8	8	105	8	13	69
53	10	20	113	12	16	61
54	8	11	109	12	15	60
55	13	22	126	8	12	68
56	19	16	142	11	11	96
57	9	19	145	5	19	97
58	17	23	155	5	17	93
59	14	28	182	9	14	96

如果在年龄 x 有 d_x 个个体死亡, 对对数似然函数的贡献 (这里假设死亡数目服从二项分布) 为

$$d_x \ln q_x + (1000 - d_x)\ln(1 - q_x).$$

似然函数在 $B = 0.000\ 243, c = 1.008\ 66, \beta_1 = 0.110$ 和 $\beta_2 = -0.014\ 4$ 处取最大值. 对于女性, 将对预期寿命 (出生) 乘因子 $\exp(0.110) = 1.116$. 血压增高 25 将使预期生存时间降低 $1 - \exp[-25(0.014\ 4)] = 0.302$ 或者 30.2%. □

习题

12.101 假设第 10 章数据集 D2 的 40 个观测来自于 4 种类型的保单持有者. 观测 $1, 5, \cdots$ 为男性吸烟者, 观测 $2, 6, \cdots$ 为男性非吸烟者, 观测 $3, 7, \cdots$ 为女性吸烟者, 观测 $4, 8, \cdots$ 为女性非吸烟者. 建立退保时间模型, 然后用这个模型估计上述 4 种情形的第一年退保概率. 根据下面要求分别构造模型.

(a) 使用 4 种不同的 Nelson-Åalen 估计, 4 组分开建模.

(b) 使用比例风险模型, 其中基准分布为指数分布.

(c) 使用比例风险模型, 且基准分布为经验分布.

12.102* 已知罢工持续时间服从 Cox 比例风险模型, 基准分布为指数分布. 唯一使用的变量是行业生产指数. 当指数值为 10 时, 罢工的持续时间的均值为 0.206 0 年. 当指数值为 25 时, 持续时间的中位数为 0.041 1 年. 试计算指数值为 5 时, 罢工持续时间大于一年的概率.

12.103* 某 Cox 比例风险模型, 男性 $z_1 = 1$, 女性 $z_1 = 0$; 成年人 $z_2 = 1$, 儿童 $z_2 = 0$. 系数的最大似然估计为 $\hat{\beta}_1 = 0.25$ 和 $\hat{\beta}_2 = -0.45$. 估计量的协方差矩阵为

$$\begin{bmatrix} 0.36 & 0.10 \\ 0.10 & 0.20 \end{bmatrix}.$$

试给出 $\beta_1 - \beta_2$ 的 95%线性置信区间, 然后由该结果构造男性儿童对女性成人的相对风险的置信区间.

12.104* 现有 4 个被保险人从出生到死亡的观测. A 组 2 个在时刻 1 和 9 死亡, B 组 2 个在时刻 2 和 4 死亡. 比例风险模型: B 组 $z_1 = 1$, A 组 $z_1 = 0$. 令 $b = \hat{\beta}_1$, 估计 A 组的某成员在时刻 3 的累积风险率.

12.105* 某 Cox 比例风险模型有 3 个协变量. 第一个死亡个体的 z_1, z_2, z_3 值为 $1, 0, 0$. 第二个死亡个体的值为 $0, 1, 0$, 第三个为 $0, 0, 1$. 试给出偏似然函数 ($\beta_1, \beta_2, \beta_3$ 的函数).

12.106 只考虑建筑材料, 重新考虑例 12.72.

12.107 使用以 Gompertz 为基准分布的比例风险模型, 重新考虑例 12.74.

12.108 使用以 gamma 为基准分布的加速失效模型, 重新考虑例 12.74.

第13章 模型选择

13.1 引 言

在数据建模过程结束时一定要从众多的模型中选择一个"优胜者". 尽管建模者可以从模型的先天条件、现实的局限性和可能会产生的误解几个方面摆脱自己对最终模型适用性的责任, 这么做是合理的, 而且常常是必不可少的, 但是通常还需要对最终的模型给出承诺. 本章将讨论模型评估和模型比较的几种不同方法. 但是也要清楚, 无论选了哪个模型, 它都是对实际情况的一种近似, 正如一位建模者的格言①:

所有的模型都是有误的, 但有些模型是可用的.

因此, 我们的工作目标是找到一个能够解决问题的足够好的模型, 而这里的主要挑战是足够好的定义将取决于具体的应用. 建模的另一个要素是, 只有对要解决的问题有足够的理解才能引导我们找到答案. 下面引用 John Tukey([131]13~14 页) 的一段话作为总结:

为正确的问题给出近似的答案要比为一个错误的问题给出准确答案有意义得多, 尽管前者是含糊的, 而后者往往可以做得非常精确.

本章将考虑一个特定的建模策略, 我们希望找到一种适用于任何概率模型的统一方法. 但这必然导致, 在任何具体的模型背景下, 都有可能存在更优 (更可靠或更精确) 的方法. 例如, 最大似然方法对大多数情形都是一个很好的估计, 但它对某些分布也不见得是最优② 的方法. 回顾各类相关的文献可以找到各种具体分布的最优估计方法, 这里不再赘述. 类似地, 这里使用的许多假设检验都给出了近似的结果. 对某个具体的情形, 可以得到更优的近似, 甚至可能是精确的结果, 本文也不介绍这些结果. 我们的目的是总结一种在大多数情况下都可以给出合理答案并能够得到广泛应用的方法.

本章内容假设读者已经具备统计假设检验的基础知识, 第 9 章对这些知识进行了复习. 以下几节的内容介绍了模型评价和模型选择的各种工具. 每种工具都有其自身的优势和弱点, 进而用不同的工具进行选择, 可能得到不同的模型. 这使得建模过程在作为一门科学的同时, 也是一门艺术. 在实际应用中, 建模的出发点将

---

① 通常认为它源自 George Box.

② "最优" 的定义有很多种, 如果用无偏性和方差最小作为 "最优" 的定义, 由 Cramér-Rao 下界以及定理 12.13 知, 最大似然估计是渐近最优的.

促使分析者对某一种工具更偏好.

13.2 数据和模型的表示

本章将要介绍的所有方法都试图将备选的模型与实际数据进行比较, 或者与另一个模型进行比较. 这个备选模型可以通过密度或分布函数表示, 也可以是由其决定的函数, 例如带上限的期望值函数或者平均剩余生存时间函数. 可以用经验分布函数或者直方图来表示数据. 如果是完整的个体数据, 这些图形都很容易得到. 如果为分组或者截断、删失的数据, 将有一定的困难. 本章涉及的情形仅仅是所有数据都在同一点 (可以是零点) 截断且在同一点 (可以是无穷) 删失的情况, 文献 [109] 对这种情形有所推广, 即存在不同截断点或删失的情况①. 应当注意到的是, 这样的表示方法仅仅在连续模型中有用. 对于离散的模型, 很少使用删失、截断以及分组处理. 数据可以很容易地由其在每个可能的观测值处的相对频率或者累积频率表示.

为了更好的表示数据, 经验分布函数将用于个体数据, 直方图用于分组数据.

为了将模型和截断的数据作比较, 我们注意到经验分布的起始点就是截断点, 实际上表示的是条件概率值 (即这种分布函数和密度函数体现的是在给定观测值超过截断点的条件下的概率). 为了和经验值作比较, 模型也必须是截断的. 设数据集的截断点是 t, 修正后的函数为

$$F^*(x) = \begin{cases} 0, & x < t, \\ \dfrac{F(x) - F(t)}{1 - F(t)}, & x \geqslant t, \end{cases}$$

$$f^*(x) = \begin{cases} 0, & x < t, \\ \dfrac{f(x)}{1 - F(t)}, & x \geqslant t. \end{cases}$$

本章中, 如果分布函数或者密度函数的脚标为样本容量, 则表示它是经验模型 (由 Kaplan-Meier、Nelson- Åalen、卵形线等方法得来) 下的分布函数或密度函数; 如果没有多余的装饰记号或者用星号 (*) 标注, 则表示估计的参数模型. 这里没有使用任何记号来表示真实的、内在的分布, 因为它不仅是未知而且是不可知的.

13.3 密度函数与分布函数的图像比较

考察模型与数据匹配程度的最直接方法是分别对密度函数和分布函数作图.

① 因为 Kaplan-Meier 估计可以用来表示在不同点截断和删失的数据, 构造模型和数据的图像并非难事. 主要的难点在于这种情形下的假设检验一般化问题.

例 13.1 考虑数据集 B 和数据集 C. 在此例及下面的例子中, 将数据集 B 中的 15 743 点用 3 476 代替 (这样处理只是为了让作出的图形能够放在一页). 表 13-1 和表 13-2 为这两个数据集的最新结果. 数据集 B 在 50 处截断, 数据集 C 在 7 500 处截断, 试用指数模型拟合这两组数据, 估计指数模型的参数. 选择适当的函数作出图形并评价该模型的拟合优度. 若数据集 B 在 1 000 处删失 (没有截断), 结果如何?

表 13-1 最大值调整后的数据集 B

27	82	115	126	155	161	243	294	340	384
457	680	855	877	974	1 193	1 340	1 884	2 558	3 476

表 13-2 数据集 C

赔付额范围	赔付笔数
0~7 500	99
7 500~17 500	42
17 500~32 500	29
32 500~67 500	28
67 500~125 000	17
125 000~300 000	9
300 000 以上	3

解 数据集 B 共有 19 个观测值 (第一个观测值被截断). 似然函数的每一项都形如 $f(82)/[1-F(50)]$. 解得指数分布参数的最大似然估计为 $\hat{\theta}=802.32$, 从 50 开始的经验分布函数在每个数据点跳跃 1/19. 若在 50 截断, 分布函数为

$$F^*(x)=\frac{1-\mathrm{e}^{-x/802.32}-(1-\mathrm{e}^{-50/802.32})}{1-(1-\mathrm{e}^{-50/802.32})}=1-\mathrm{e}^{-(x-50)/802.32}.$$

图 13-1 给出了这两个函数的图像.

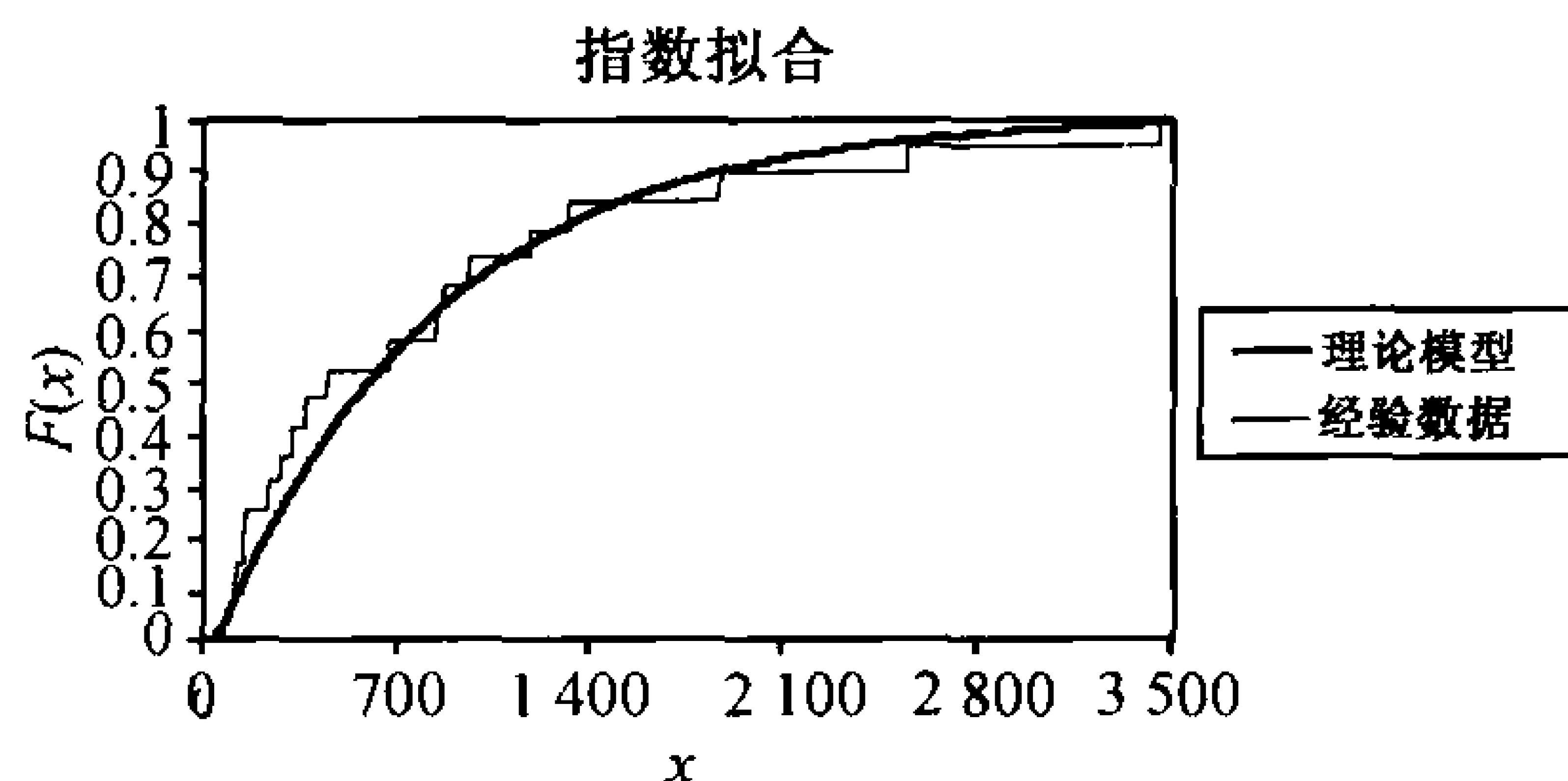


图 13-1 数据集 B 在 50 处截断时模型和数据的积累分布函数图像

这个拟合效果并不令人满意, 因为在 x 取较小值处模型低估了分布函数, 而在 x 取较大值处模型高估了分布函数. 这种不合适是因为它意味着模型低估了尾概

率.

对于数据集 C 似然函数要考虑截断情况, 例如, 第一个区间对似然函数的贡献是

$$\left[\frac{F(17\ 500)-F(7\ 500)}{1-F(7\ 500)}\right]^{42}.$$

最大似然估计是 $\hat{\theta}=44\ 253$, 直方图第一段的高度是

$$\frac{42}{128(17\ 500-7\ 500)}=0.000\ 032\ 8.$$

最后一段柱状图的范围是从 125 000 到 300 000(从 300 000 到无穷范围的柱状图是无法构造的). 密度函数必须在 7 500 处截断, 因此变成了

$$\begin{aligned} f^*(x) &= \frac{f(x)}{1-F(7\ 500)}=\frac{44\ 253^{-1}\mathrm{e}^{-x/44\ 253}}{1-(1-\mathrm{e}^{-7\ 500/44\ 253})} \\ &= \frac{\mathrm{e}^{-(x-7\ 500)/44\ 253}}{44\ 253}, \quad x>7\ 500. \end{aligned}$$

密度函数的拟合和直方图如图 13-2 所示.

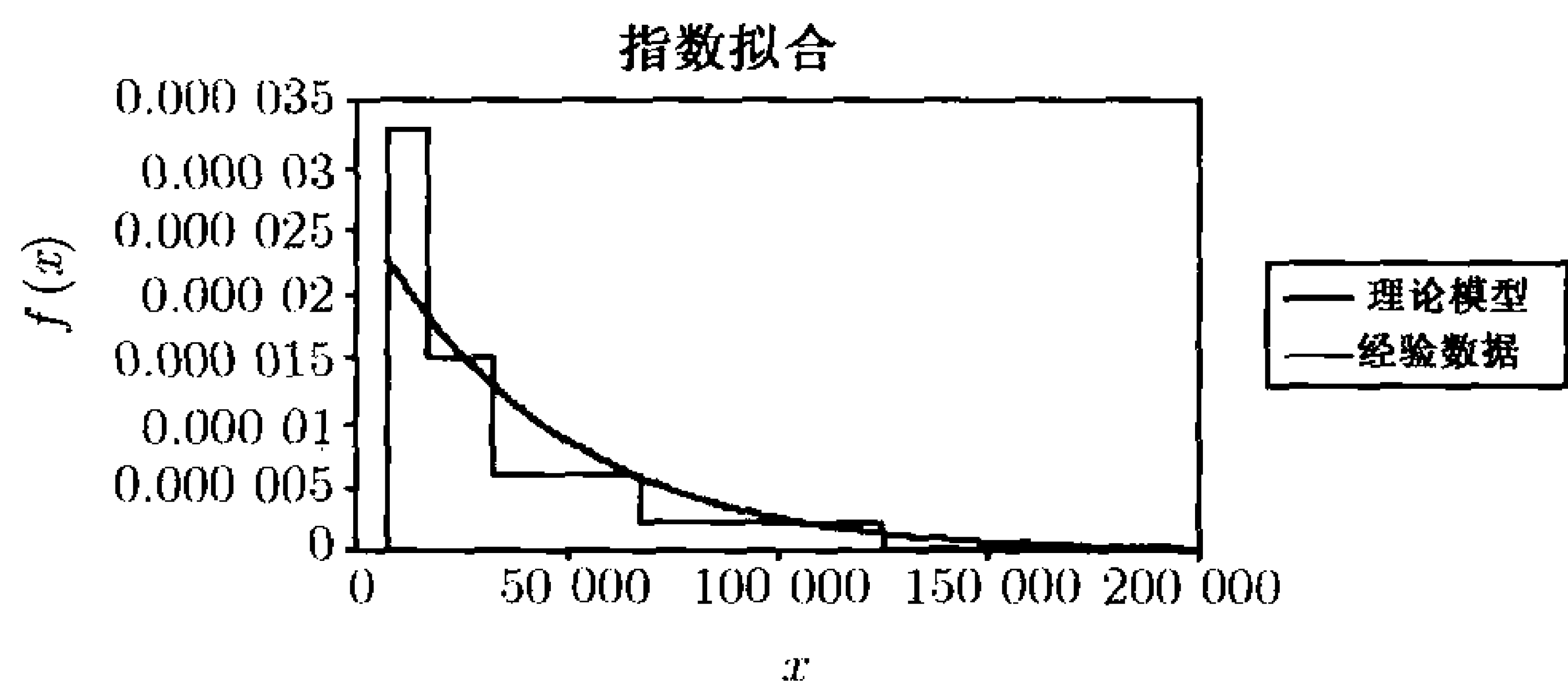


图 13-2　数据集 C 在 7 500 处截断时模型和数据的密度图像

在自变量比较小时, 指数模型低估了密度函数; 从图形中很难看出 125 000 以后两条曲线的差异情况.

当用上界 1 000 修正数据集 B 后, 最大似然估计为 $\hat{\theta}=718.00$, 经验分布函数的图形必须在 1 000 处截止, 见图 13-3.

又一次看出, 指数模型的拟合效果不佳. □

当模型的分布函数和经验分布函数很接近时, 很难从图像上分辨出细微的差别. 但也有许多办法可以将这些细微的差别放大, 这里介绍其中两种. 第一种是直接画出两个函数差值的图像. 也就是说, 如果 $F_n(x)$ 和 $F^*(x)$ 分别表示经验分布函数和由模型得到的分布函数, 画出 $D(x)=F_n(x)-F^*(x)$ 的图像即可.

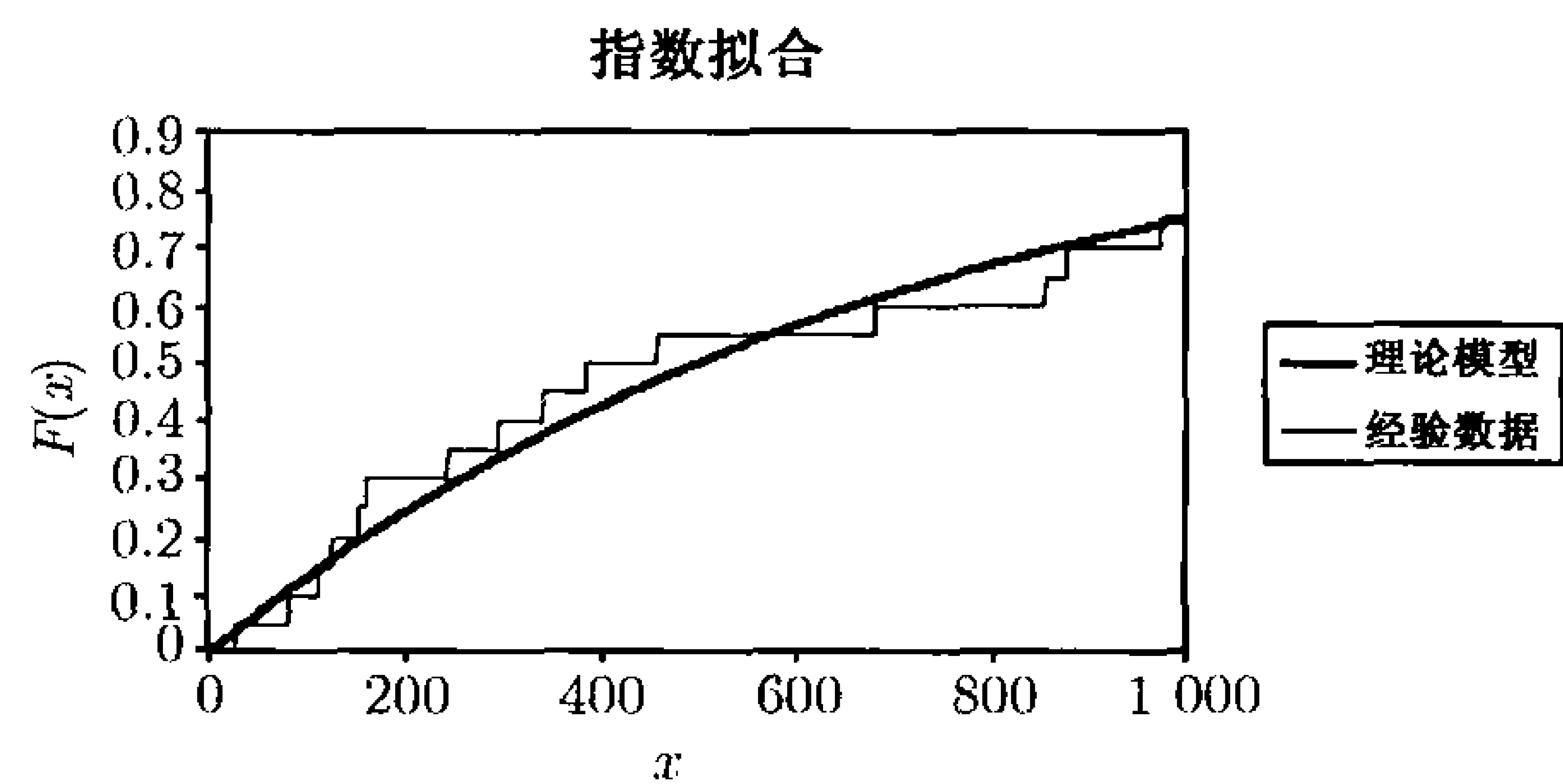


图 13-3 数据集 B 在 1 000 处删失时模型和数据的累积分布函数图像

例 13.2 试给出例 13.1 中 $D(x)$ 的图像.

解 对于在 50 处截断的数据集 B, 图像如图 13-4 所示. 该模型较差的拟合效果在这个图像中得到了充分的表现.

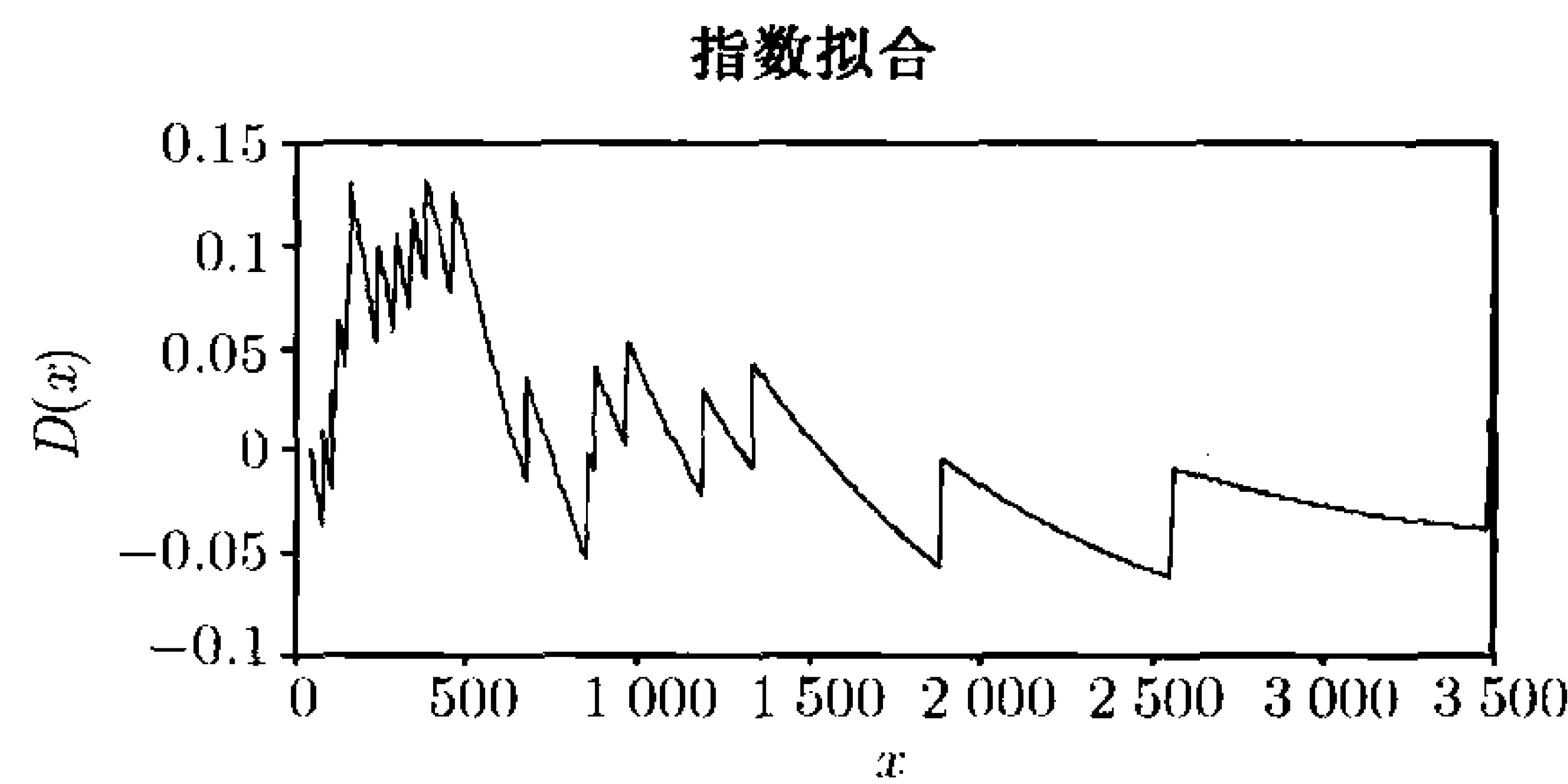


图 13-4 数据集 B 在 50 处截尾时模型和数据的 $D(x)$ 函数图像

对于分组的数据, 没有与之对应的 $D(x)$ 图形. 对于在 1 000 处删失的数据集 B, 图像也在 1 000 处截止, 见图 13-5. 这个图像又一次体现了该模型较差的拟合效果. □

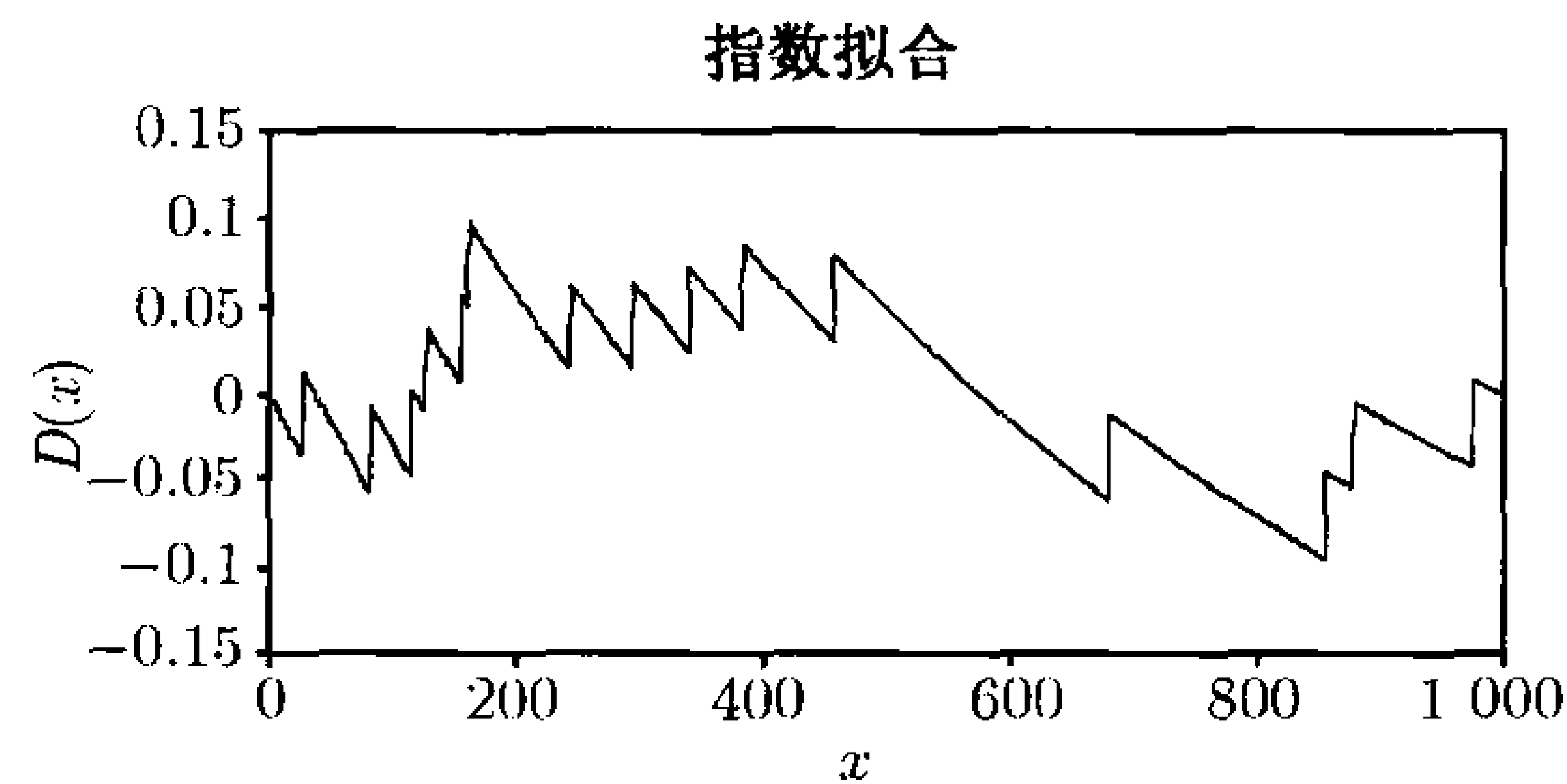


图 13-5 数据集 B 在 1 000 处删失时模型和数据的 $D(x)$ 函数图像

另一种显示拟合误差的方法是 $p-p$ 图, 也称概率图 (probability plot). 首先将观测值排序 $x_1 \leqslant \cdots \leqslant x_n$, 再对每个值构造坐标 $(F_n(x_j), F^*(x_j))$, 最后将每个坐标对应的点画在 $(F_n(x), F_n^*(x))$ 的平面上[①]. 如果模型拟合得很好, 描出的各个点都应该在从 (0, 0) 到 (1, 1) 的直线附近. 但是, 在这种情况下, 必须对经验分布函数的定义有所修改. 因为可以证明, $F_n(x_j)$ 的期望值为 $j/(n+1)$, 进而经验分布函数在该点的值也应当是这个值而非通常的取值 j/n. 对两个相同的观测值可以直接标为两个点 (它们有相同的 "y" 坐标但 "x" 坐标不同), 也可以取 "x" 坐标的平均值只画一个点.

例 13.3 (续例 13.2), 画出 $p-p$ 图.

解 当数据集 B 在 50 截断时, $n=19$, 以观测值 $x=82$ 为例, 经验分布值为 $F_n(82)=1/20=0.05$, 另一个坐标值是

$$F^*(82)=1-\mathrm{e}^{-(82-50)/802.32}=0.039\ 1.$$

这就得到 $p-p$ 图中的一个点 (0.05, 0.039 1). 类似地, 可以得到所有的点, 其图形如图 13-6 所示.

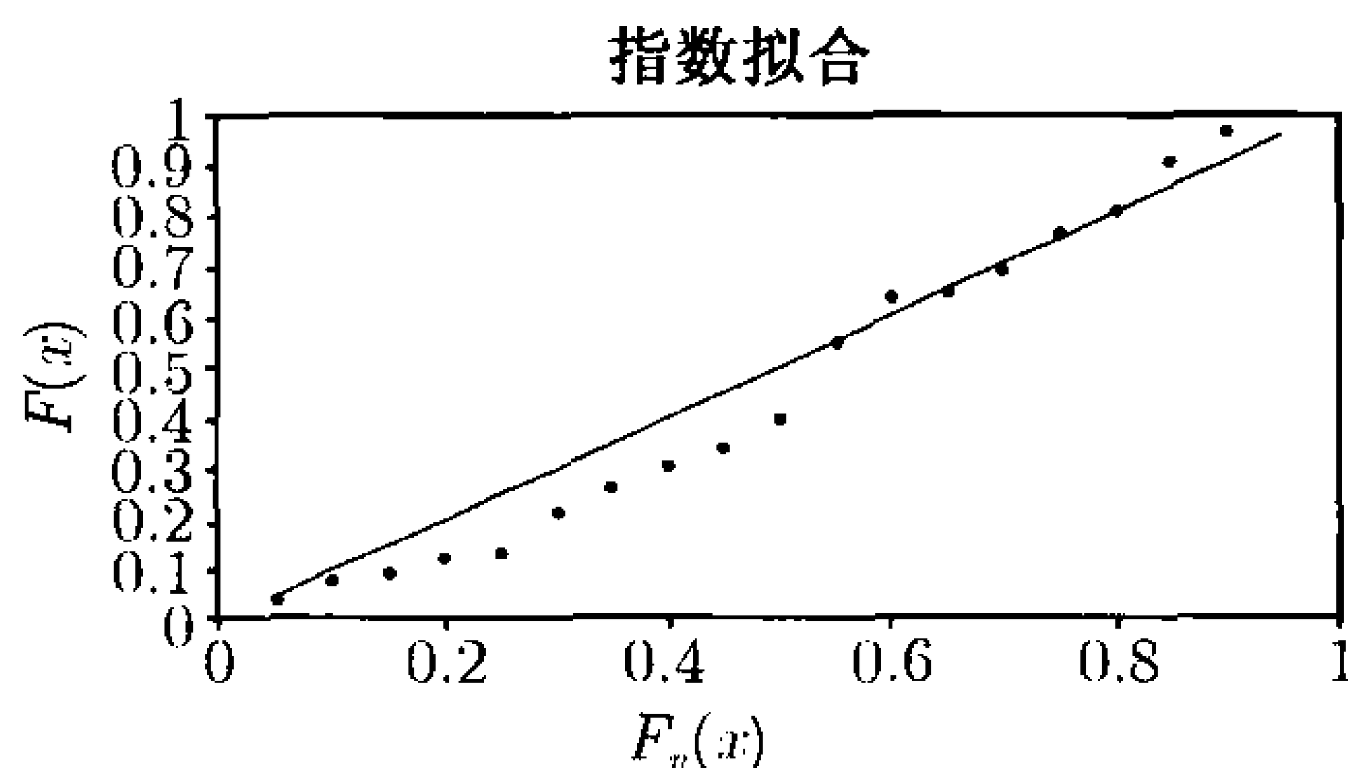


图 13-6 数据集 B 在 50 处截断时的 $p-p$ 图

在图的左下方, 指数模型对小观测值的概率明显小于数据给出的. 对数据集 B 在 1 000 删失的情况, 也可由类似的办法得到 $p-p$ 图, 见图 13-7.

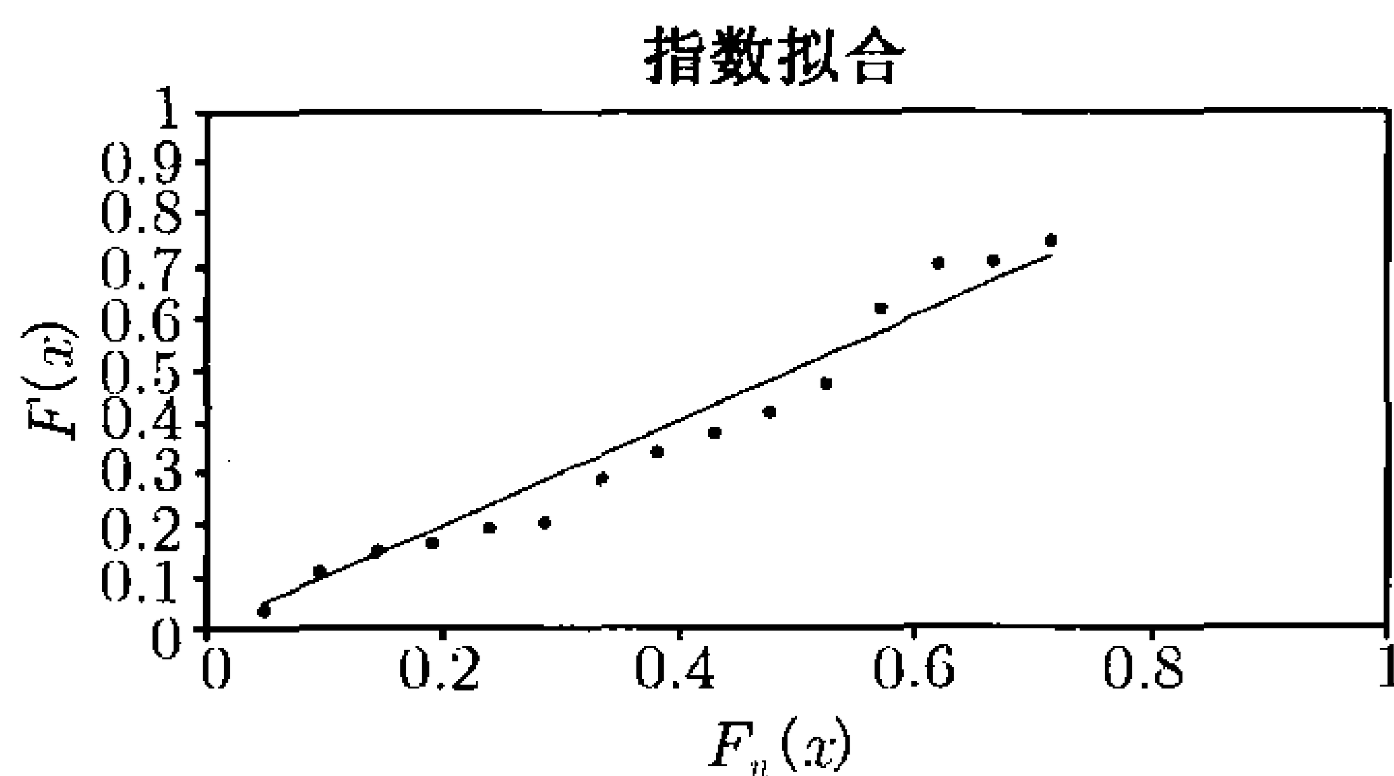


图 13-7 数据集 B 在 1 000 处删失时的 $p-p$ 图

① 本书的第一版误称图形为 "$q-q$ 图". $q-q$ 图是另外一种常用的图像, 这里不作介绍.

这个图像在 0.75 左右截止了, 因为这是删失点 1 000 左边的最大概率, 没有更高概率的经验值. 指数模型又一次低估于经验值. □

习题

13.1 用 Weibull 模型代替指数模型, 重新计算例 13.1.

13.2 用 Weibull 模型, 重新计算例 13.2.

13.3 用 Weibull 模型, 重新计算例 13.3.

13.4 假 设 检 验

有时图像的确可以胜过很多的文字, 但有时候用数学论证来代替图形传递给我们的印象才是最佳的选择. 统计假设检验就是上述 "论证" 的一种:

H_0: 数据来自于某个给定的总体;

H_1: 数据并非来自于这个总体.

检验统计量往往是对模型的分布函数和经验分布函数之间接近程度的一个度量. 如果在零假设中完全指明了模型 (比如, 均值为 100 的指数分布), 检验的临界值就很容易得到. 但是, 更常见的情况是, 零假设仅仅指明了模型的名称但没有给出参数. 如果模型的参数是通过数据估计出来的, 这时的检验统计量比先前给定模型参数的情况更小. 这是因为参数估计时就是尽量使得分布函数与实际数据接近. 这样, 检验就变成了近似. 由于统计量较大时为拒绝零假设, 这种近似在增加犯第二类错误概率的同时降低了犯第一类错误的概率[①]. 在精算的模型中, 这是一个可以接受的代价.

一种避免近似状况出现的方式是将样本随机地分为两部分, 其中的一半进行参数估计, 用另一半进行假设检验. 一旦模型选定了, 所有的数据都可以用来重新对参数进行估计.

13.4.1 Kolmogorov-Smirnov 检验

令 t 为左截断点 (如果没有截断则 $t=0$), u 为右删失点 (如果无删失则 $u=\infty$). 这时的检验统计量为

$$D=\max_{t\leqslant x\leqslant u}|F_n(x)-F^*(x)|.$$

为了确保阶梯函数 $F_n(x)$ 有定义, 这个统计量只适用于个体数据. 另外, 这里假设模型的分布函数 $F^*(x)$ 在对应的区间上是连续的.

例 13.4 计算例 13.1 的 D 值.

① 在本书介绍的检验中, 只有卡方检验可以对此情形作出修正, 这是由其内在的原理决定的. 对其他检验的修正方法也将陆续提出, 但这里不作介绍.

解 表 13-3 给出了计算需要的值. 由于经验分布函数在每个数据点跳跃, 因此既要将模型分布函数与跳跃前的值比较又要与跳跃后的值比较. 表中将跳跃前的函数值记为 $F_n(x-)$. D 的最大值是 0.134 0.

表 13-3 例 13.4 中 D 的计算

x	$F^*(x)$	$F_n(x-)$	$F_n(x)$	最大差距
82	0.039 1	0.000 0	0.052 6	0.039 1
115	0.077 8	0.052 6	0.105 3	0.027 5
126	0.090 4	0.105 3	0.157 9	0.067 5
155	0.122 7	0.157 9	0.210 5	0.087 8
161	0.129 2	0.210 5	0.263 2	0.134 0
243	0.213 8	0.263 2	0.315 8	0.102 0
294	0.262 2	0.315 8	0.368 4	0.106 2
340	0.303 3	0.368 4	0.421 1	0.117 8
384	0.340 5	0.421 1	0.473 7	0.133 2
457	0.397 9	0.473 7	0.526 3	0.128 4
680	0.544 0	0.526 3	0.578 9	0.034 9
855	0.633 3	0.578 9	0.631 6	0.054 4
877	0.643 3	0.631 6	0.684 2	0.040 9
974	0.683 9	0.684 2	0.736 8	0.052 9
1 193	0.759 4	0.736 8	0.789 5	0.030 1
1 340	0.799 7	0.789 5	0.842 1	0.042 4
1 884	0.898 3	0.842 1	0.894 7	0.056 2
2 558	0.956 1	0.894 7	0.947 4	0.061 4
3 476	0.986 0	0.947 4	1.000 0	0.038 6

数据集 B 在点 1 000 删失, 20 个观测值中有 15 个未删失. 表 13-4 显示了计算的步骤. D 的最大值是 0.099 1. □

表 13-4 例 13.4 中删失数据的情况下 D 的计算

x	$F^*(x)$	$F_n(x-)$	$F_n(x)$	最大差距
27	0.036 9	0.00	0.05	0.036 9
82	0.107 9	0.05	0.10	0.057 9
115	0.148 0	0.10	0.15	0.048 0
126	0.161 0	0.15	0.20	0.039 0
155	0.194 2	0.20	0.25	0.055 8
161	0.200 9	0.25	0.30	0.099 1
243	0.287 1	0.30	0.35	0.062 9
294	0.336 0	0.35	0.40	0.064 0
340	0.377 2	0.40	0.45	0.072 8
384	0.414 2	0.45	0.50	0.085 8
457	0.470 9	0.50	0.55	0.079 1

(续)

x	$F^*(x)$	$F_n(x-)$	$F_n(x)$	最大差距
680	0.612 1	0.55	0.60	0.062 1
855	0.696 0	0.60	0.65	0.096 0
877	0.705 2	0.65	0.70	0.055 2
974	0.742 5	0.70	0.75	0.042 5
1 000	0.751 6	0.75	0.75	0.001 6

现在余下的任务是临界值的计算. 该检验常用的临界值计算方法是, 当 $\alpha = 0.10$ 时临界值取为 $1.22/\sqrt{n}$, 当 $\alpha = 0.05$ 时临界值取为 $1.36/\sqrt{n}$, 当 $\alpha = 0.01$ 时临界值取为 $1.63/\sqrt{n}$. 如果 $u < \infty$, 临界值还应当取得更小, 因为此时两个函数值的差距变得更大的机会更小. 有一些基于这个现象修正方法的文献论述 (例如 [125], 其中还包含具体的零假设的分布模型临界值表), [109] 中也提供了一种修正方法, 这里不作介绍.

例 13.5 完成例 13.4 的 Kolmogorov-Smirnov 检验.

解 数据集 B 在 50 处截断, 样本容量为 19. 显著性水平为 5%的临界值是 $1.36/\sqrt{19} = 0.312\ 0$, 由于 0.134 0<0.312 0, 因此不能拒绝零假设, 并且应认为指数分布似乎是一个正确的模型. 尽管看起来指数模型并不适用于这个总体, 但样本容量太小从而无法得出否定指数模型的结论. 数据集 B 在 1 000 删失时, 样本容量为 20, 进而临界值是 $1.36/\sqrt{20} = 0.304\ 1$, 指数模型仍然显得可以接受. □

不管是上述检验还是以下要介绍的 Anderson-Darling 检验, 临界值都只是在零假设为完全明确的分布模型的情况下才是正确的. 如果数据集本身对零假设的参数估计有贡献 (就如例题中那样), 正确的临界值应当稍小一些. 在这两种检验中, 临界值的变化都依赖于特定的假设分布, 甚至还可能依赖于特定的参数真实值. 17.2.4 节将介绍如何在这种情形下使用随机模拟方法.

13.4.2 Anderson-Darling 检验

这种检验方法和 Kolmogorov-Smirnov 检验类似, 只是采用了另一种度量两个分布函数之间差距的方法. 检验量统计为

$$A^2 = n\int_t^u \frac{[F_n(x) - F^*(x)]^2}{F^*(x)[1 - F^*(x)]} f^*(x)\mathrm{d}x.$$

这是对经验分布函数与模型分布函数的平方误差的加权平均. 注意到当 x 接近 t 或者 u 的时候, 由于分母的某一个因子将变得很小, 权重会变得非常大. 这个检验统计量显得更加重视尾部的拟合优度而非分布函数的中间部分. 用这个公式进行计算看起来并不是一件容易的事情. 但是对于个体数据 (说明这个检验也不适用于

分组数据), 这个积分可以被化简为

$$A^2 = -nF^*(u) + n\sum_{j=0}^{k}[1 - F_n(y_j)]^2\{\ln[1 - F^*(y_j)] - \ln[1 - F^*(y_{j+1})]\}$$

$$+ n\sum_{j=1}^{k} F_n(y_j)^2[\ln F^*(y_{j+1}) - \ln F^*(y_j)],$$

其中无重复的未删失数据点为 $t = y_0 < y_1 < \cdots < y_k < y_{k+1} = u$. 请注意当 $u = \infty$ 时第一个求和式的最后一项是零 [若直接计算会产生 ln(0) 的问题]. 显著性水平为 10%、5%和 1%的临界值分别是 1.933、2.492 和 3.875. 和 Kolmogorov-Smirnov 检验一样, 如果 $u < \infty$, 这些临界值还应当更小一些.

例 13.6 (续例 13.5) 考虑 Anderson-Darling 检验.

解 数据集 B 在 50 截断时共有 19 个数据点, 计算过程如表 13-5 所示, 其中 "求和项" 一栏指的是公式中两个求和式中对应的项之和. 总和为 1.022 6, 检验统计量为 $-19(1) + 19(1.022\,6) = 0.429\,2$. 由于检验统计量小于临界值 2.492, 因此指数模型显得有一定的合理性.

表 13-5 例 13.6 的 Anderson-Darling 检验

j	y_j	$F^*(x)$	$F_n(x)$	求和项
0	50	0.000 0	0.000 0	0.039 9
1	82	0.039 1	0.052 6	0.038 8
2	115	0.077 8	0.105 3	0.012 6
3	126	0.090 4	0.157 9	0.033 2
4	155	0.122 7	0.210 5	0.007 0
5	161	0.129 2	0.263 2	0.090 4
6	243	0.213 8	0.315 8	0.050 1
7	294	0.262 2	0.368 4	0.042 6
8	340	0.303 3	0.421 1	0.038 9
9	384	0.340 5	0.473 7	0.060 1
10	457	0.397 9	0.526 3	0.149 0
11	680	0.544 0	0.578 9	0.089 7
12	855	0.633 3	0.631 6	0.009 9
13	877	0.643 3	0.684 2	0.040 7
14	974	0.683 9	0.736 8	0.075 8
15	1 193	0.759 4	0.789 5	0.040 3
16	1 340	0.799 7	0.842 1	0.099 4
17	1 884	0.898 3	0.894 7	0.059 2
18	2 558	0.956 1	0.947 4	0.030 8
19	3 476	0.986 0	1.000 0	0.014 1
20	∞	1.000 0	1.000 0	

数据集 B 在 1 000 删失时的结果如表 13-6 所示. 总和为 0.760 2, 检验统计量为 $-20(0.751\,6)+20(0.760\,2)=0.171\,3$. 由于检验统计量未超过临界值 2.492, 指数模型仍然可以接受. □

表 13-6 例 13.6 中删失数据 Anderson-Darling 检验的计算

j	y_j	$F^*(x)$	$F_n^*(x)$	求和项
0	0	0.000 0	0.00	0.037 6
1	27	0.036 9	0.05	0.071 8
2	82	0.107 9	0.10	0.040 4
3	115	0.148 0	0.15	0.013 0
4	126	0.161 0	0.20	0.033 4
5	155	0.194 2	0.25	0.006 8
6	161	0.200 9	0.30	0.088 1
7	243	0.287 1	0.35	0.049 3
8	294	0.336 0	0.40	0.041 6
9	340	0.377 2	0.45	0.037 5
10	384	0.414 2	0.50	0.057 5
11	457	0.470 9	0.55	0.142 3
12	680	0.612 1	0.60	0.085 2
13	855	0.696 0	0.65	0.009 3
14	877	0.705 2	0.70	0.037 4
15	974	0.742 5	0.75	0.009 2
16	1 000	0.751 6	0.75	

13.4.3 卡方 (χ^2) 拟合优度检验

与前文介绍的两种检验方法不同, 这个检验方法有更大的灵活性. 首先, 任意选定 $k-1$ 个值 $t=c_0<c_1<\cdots<c_k=\infty$, 并记 $\hat{p}_j=F^*(c_j)-F^*(c_{j-1})$ 为观测值落在 c_{j-1} 到 c_j 区间中的概率. 类似地, 记 $p_{nj}=F_n(c_j)-F_n(c_{j-1})$ 为经验分布计算的上述概率. 此时检验统计量为

$$\chi^2=\sum_{j=1}^{k}\frac{n(\hat{p}_j-p_{nj})^2}{\hat{p}_j},$$

其中 n 为样本量. 然后再考虑该公式的另一种表示法: 令 $E_j=n\hat{p}_j$ 为区间中观测值个数的期望值 (认为假设的模型是正确的) 并令 $O_j=np_{nj}$ 为区间中的实际观测次数. 此时有

$$\chi^2=\sum_{j=1}^{k}\frac{(E_j-O_j)^2}{E_j}.$$

该检验的临界值由相应的卡方分布决定, 该卡方分布的自由度应该是求和式中的项数 (k) 减去 1 再减去待估参数的个数. 人们设计了一系列规则来描述这个检

验何时可以达到合理的精度. 这些方法都是围绕 $E_j = n\hat{p}_j$ 进行的. 最保守的看法是 E_j 的值应当不小于 5, 不过一些作者也提出, 即使有的值低到 1, 仍是可以接受的. 公认的一点是, 当这些 E_j 的值大致相等时, 检验方法的效果是最好的. 虽然可以将相邻的组并在一起使得 E_j 变大, 但是如果数据已经分组, 除了采用给定的分组, 选择余地却很小. 为了对个体数据进行该检验, 应将数据适当分组①.

例 13.7 接上例, 对指数分布进行卡方拟合优度检验.

解 3 个数据集都可以进行这种检验. 数据集 B 在 50 截断时, 设分界点为 50、150、250、500、1 000、2 000 以及无穷大. 表 13-7 为计算过程, 总 χ^2 值为 1.403 4. 在 5%的显著性水平下, 4 个自由度 (6 组减 1 再减去 1 个待估参数) 的临界值为 9.487 7(可由 Excel® 中的函数 CHIINV(0.05, 4) 求得), p 值为 0.843 6[由 CHIDIST (1.403 4, 4) 得到]. 由此可见指数模型拟合效果较好.

表 13-7 在 50 处截断的数据集 B

范 围	$\hat{p}$	期望值	观测值	χ^2
50~150	0.117 2	2.227	3	0.268 7
150~250	0.103 5	1.966	3	0.544 4
250~500	0.208 7	3.964	4	0.000 3
500~1 000	0.264 7	5.029	4	0.210 5
1 000~2 000	0.218 0	4.143	3	0.315 2
2 000~∞	0.088 0	1.672	2	0.064 4
总计	1	19	19	1.403 4

数据集 B 在 1 000 删失时, 第一个区间是 0~150, 最后一个区间是 1 000~∞. 与前面两种检验法相比, 卡方检验可以用删失的观测值来实现. 计算如表 13-8 所示, 总 χ^2 值为 0.595 1. 在 5%的显著性水平下, 3 个自由度 (5 组减 1 再减去 1 个待估参数) 的临界值为 7.817 4, p 值为 0.897 6, 可见指数模型拟合效果较好.

表 13-8 在 1 000 处删失的数据集 B

范 围	$\hat{p}$	期望值	观测值	χ^2
0~150	0.188 5	3.771	4	0.013 9
150~250	0.105 5	2.110	3	0.375 4
250~500	0.207 6	4.152	4	0.005 5
500~1 000	0.250 0	5.000	4	0.200 0
1 000~∞	0.248 4	4.968	5	0.000 2
总计	1	20	20	0.595 1

① Moore [95] 引用了诸多规则, 其中有以下几条: (1) 所有单元的期望频数都不低于 1, 并且至少有 80%的单元的期望频数不低于 5; (2) 在显著性水平为 1%的检验中, 每个单元观测数目的平均值至少为 4, 在显著性水平为 5%的检验中, 每个单元观测数目的平均值至少为 2; (3) 样本容量至少为 10 并至少有 3 个单元, 而且样本量的平方与单元个数的比值至少为 10.

数据集 C 已经分组, 其计算如表 13-9 所示. 检验统计量 $\chi^2 = 61.913$, 自由度为 4, 临界值为 9.488, p 值约为 10^{-12}, 因此有充分的理由认为指数模型是不合适的. 为了进行更精确的检验, 应将最后两个组合并在一起 (因为最后一个组的期望值小于 1). 从 125 000 到无穷大这个组的期望值是 8.997, 观测值为 12 并有 1.002 的贡献. 检验统计量为 16.552, 自由度为 3, p 值为 0.000 87. 检验结果仍然拒绝了指数模型. □

表 13-9 数据集 C

范　围	$\hat{p}$	期望值	观测值	χ^2
7 500～17 500	0.202 3	25.889	42	10.026
17 500～32 500	0.229 3	29.356	29	0.004
32 500～67 500	0.310 7	39.765	28	3.481
67 500～125 000	0.187 4	23.993	17	2.038
125 000～300 000	0.068 9	8.824	9	0.003
300 000～∞	0.001 3	0.172	3	46.360
总计	1	128	128	61.913

有时, 为了适应不同的情形, 可能对检验方法稍作修改. 以下这个例子说明了如何运用总频率的数据进行检验.

例 13.8 对例 12.60 的 Poisson 模型, 构造一个近似的拟合优度检验. 所用的数据如表 13-10 所示.

表 13-10 每年车险索赔额

年份	潜在风险	索赔数
1986	2 145	207
1987	2 452	227
1988	3 112	341
1989	3 458	335
1990	3 698	362
1991	3 872	359

解 假设每年的索赔数为一组 (由潜在风险给出) 独立同分布的随机变量之和. 在这种情形下, 由中心极限定理, 正态估计是恰当的. 期望数 (E_k) 为潜在风险乘以每个潜在风险单位的期望值, 方差 (V_k) 为潜在风险乘以某个潜在风险单位的方差的估计值. 进而, 检验统计量为

$$Q = \sum_k \frac{(n_k - E_k)^2}{V_k},$$

近似服从自由度等于数据点个数减去待估计参数个数的卡方分布. 期望数 $E_k =$

λe_k, 对方差也有 $V_k = \lambda e_k$, 检验统计量为

$$\begin{aligned}Q = &\frac{(207-209.61)^2}{209.61} + \frac{(227-239.61)^2}{239.61} + \frac{(341-304.11)^2}{304.11} \\ &+ \frac{(335-337.92)^2}{337.92} + \frac{(362-361.37)^2}{361.37} + \frac{(359-378.38)^2}{378.38} \\ = &6.19.\end{aligned}$$

对 5 个自由度的情况, 5%的临界值为 11.07, 因此接受 Poisson 假设. □

关于这些检验法, 下面这个问题特别重要, 假设当样本容量加倍时但样本点的取值并没有太大变化 (想象每个数据点都重复出现两次). 在 Kolmogorov-Smirnov 检验中检验统计量不变, 但临界值会变小. 在 Anderson-Darling 和卡方检验中, 检验统计量会加倍但临界值不会改变. 结果是, 对于较大容量的样本, 零假设 (即备选模型) 更容易被拒绝. 这并不奇怪, 我们知道, 零假设本身其实是错误的 (由几个参数决定的某个分布函数就能够合理解释观测值表现的各种复杂现象, 这种解释成立的可能性极小), 但只有在样本容量足够大的时候, 才会有可以令人信服的证据来说明这一点. 在使用这些检验方法时必须牢记, 尽管所有的模型都是有误的, 但有些模型是可用的.

13.4.4 似然比检验

相对 "总体是否服从 A 分布?" 的另一种问题是 "总体更可能具有 B 分布还是 A 分布?" 后一类的正规表达方式为

H_0 : 数据来自服从 A 分布的总体,

H_1 : 数据来自服从 B 分布的总体.

为了能够进行正规的假设检验, A 分布必须是 B 分布的一种特殊情形, 比如, 指数分布相对于 gamma 分布. 以下将介绍完成这种检验的一个简单的方法.

定义 13.9 如下描述的检验称为似然比检验.首先将似然函数记为 $L(\theta)$, 用 θ_0 表示使得似然函数取到最大值的参数值, 不过只能选择零假设允许范围内的参数. 令 $L_0 = L(\theta_0)$. 用 θ_1 表示在备择假设成立的所有可能的参数值范围内变化时, 使得似然函数最大的参数, 并记 $L_1 = L(\theta_1)$. 检验统计量为 $T = 2\ln(L_1/L_0) = 2(\ln L_1 - \ln L_0)$. 如果 $T > c$, 则拒绝零假设, 其中 c 满足 $\alpha = \Pr(T > c)$, T 服从卡方分布, 其自由度等于备择假设中模型的自由参数个数减去零假设中模型的自由参数个数.

这个检验法具有一定的意义. 因为当备择假设为真时, 将使得由零假设的参数计算的似然函数值相当的小.

例 13.10 检验下述的假设: 数据集 B(用最初的观测值) 的总体均值不是 1 200. 已知总体服从 gamma 分布, 显著性水平取为 5%, 采用似然比检验, 并计算 p 值.

解 假设为

$$H_0: \mu = 1\ 200\text{的 gamma 分布},$$

$$H_1: \mu \neq 1\ 200\text{的 gamma 分布}.$$

由前面的结果, 最大似然估计为 $\hat{\alpha} = 0.556\ 16, \hat{\theta} = 2\ 561.1$, 对数似然函数的最大值为 $\ln L_1 = -162.293$. 下一步, 考虑在 α 和 θ 的取值满足 $\alpha\theta = 1\ 200$ 的前提下使似然函数最大化. 这意味着 α 可以取任意的正实数, 但是 θ 必须满足:$\theta = 1\ 200/\theta$, 在零假设下, 只有一个自由参数, 似然函数在 $\hat{\alpha} = 0.549\ 55, \hat{\theta} = 2\ 183.6$ 时取到最大值, 对数似然函数的最大值为 $\ln L_0 = 162.466$. 检验统计量为 $T = 2(-162.293 + 162.466) = 0.346$. 而对于一个自由度的卡方分布, 临界值为 3.841 5, 由于 0.346<3.841 5, 不拒绝零假设. 一个自由度的卡方分布的随机变量超过 0.346 的概率是 0.556, 该 p 值也不支持备择假设. □

例 13.11 (续例 4.42) 已知 $(a, b, 0)$ 类型分布族不足以描述原题中的数据, 请提出一个合理的模型.

解 选择 13 种不同的模型对数据进行拟合. 结果显示, 对于卡方拟合优度检验, 其中有 6 个模型的 p 值在 0.01 以上, 这 6 个模型的信息如表 13-11 所示. 似然比检验可以发现, 负对数似然函数最小的三参数模型 (Poisson-ETNB 模型) 与两参数的 Poisson- 逆高斯模型相比, 并没有表现出明显的优势, 因此后者看上去是一个极佳的选择. □

表 13-11 例 13.11 可用的六种模型

模型	参数个数	负对数似然函数	χ^2	p 值
负二项	2	5 348.04	8.77	0.012 5
零调整的对数	2	5 343.79	4.92	0.177 9
Poisson- 逆高斯	2	5 343.51	4.54	0.209 1
零调整的负二项	3	5 343.62	4.65	0.097 9
几何 - 负二项	3	5 342.70	1.96	0.375 4
Poisson-ETNB	3	5 342.51	2.75	0.252 5

例 13.12 例 12.65 中的 β_1 估计值很小. 用 beta 模型进行似然比检验, 判断年龄是否对损失量有显著的影响.

解 令 $\beta_1 = 0$, 重新估计参数 $\hat{\beta}_2 = -0.791\ 93, \hat{\alpha} = 1.031\ 18, \hat{b} = 0.742\ 49$, 对数似然函数值为 $-4.220\ 9$. 加上年龄因素, 似然函数值改进了 0.005 4, 并不显著. □

当备择分布仅比零假设具有更多的参数时, 人们总是试用这种检验. 但实际上这种情况下该检验是不合适的, 例如, 两参数的对数正态模型完全有可能比三参数的 Burr 模型的对数似然函数值要高, 这样就会产生负的检验统计量, 进而无法用卡方分布. 当零假设为备择分布的极限情形时 (而不是一种特殊情形), 这种检验仍

然适用, 但是检验统计量为混合卡方分布 (见 [120]). 无论如何, 利用该项 "检验" 在上述情形中作出决策还是有一定道理的, 不过应该清楚这并不是一种正规的假设检验. 在 13.5 节和第 14 章中, 都有进一步使用该检验进行决策的例题和习题.

习题

13.4 用 Kolmogorov-Smirnov 检验法判断 Weibull 模型是否适用于例 13.5 的数据.

13.5* 某随机变量的 5 个观测分别为 1, 2, 3, 5, 13. 零假设: $f(x)=2x^{-2}\mathrm{e}^{-2/x}, x>0$, 计算 Kolmogorov-Smirnov 检验统计量的值.

13.6* 给定以下 5 个来自同一随机样本的观测值: 0.1, 0.2, 0.5, 1.0, 1.3, 对于零假设: 总体的密度函数是 $f(x)=2(1+x)^{-3}, x>0$, 计算 Kolmogorov-Smirnov 检验统计量.

13.7 在例 13.6 中, 依据 Weibull 分布进行 Anderson-Darling 检验.

13.8 用 Weibull 模型重做例 13.7.

13.9* 对 150 名投保人, 从签订保单受益凭证开始观察, 直到其身故, 且没有删失观测值. 有 21 人在第 1 年身故, 有 27 人在第 2 年身故, 有 39 人在第 3 年, 另有 63 人在第 4 年. 考虑生存模型

$$S(t)=1-\frac{t(t+1)}{20}, \quad 0\leqslant t\leqslant 4,$$

显著性水平为 5%, 进行卡方拟合优度检验.

13.10* 365 天的索赔数记录为: 50 天没有索赔, 122 天有 1 个索赔, 101 天有 2 个索赔, 92 天有 3 个索赔, 没有 1 天发生 4 次以上的索赔. 求 Poisson 模型 λ 的最大似然估计, 并在 2.5%的显著性水平下进行卡方拟合优度检验.

13.11* 一年内每天发生的事故数分布如表 13-12 所示. 考虑如下的假设检验: 数据来自均值为 0.6 的 Poisson 分布, 请将数据分为尽可能多的组, 并保证每个组期望的观测数至少为 5, 显著水平为 5%.

表 13-12 习题 13.11 的数据

事故数目	天数
0	209
1	111
2	33
3	7
4	3
5	2

13.12 假设每个潜在风险单位都服从几何分布, 重做例 13.8, 并考虑近似卡方优度检验. 几何分布比 Poisson 分布更优吗?

13.13 基于数据集 B(使用最初的最大观测值), 判断 gamma 模型是否比指数模型更合理. 注意指数模型是 gamma 模型在 $\alpha=1$ 时的特殊情况. 可以直接利用例 12.8 中的数值.

13.14 在指数、gamma 和变形 gamma 模型中, 为数据集 C 选择一个产生其数据的总体的

模型. 例 12.9 和习题 12.21 中分别有关于前两个模型的信息.

13.15 对习题 12.96 得到的各个模型进行卡方拟合优度检验.

13.16 对习题 12.98 得到的各个模型进行卡方拟合优度检验.

13.17 对习题 12.99 得到的各个模型进行卡方拟合优度检验.

13.18 基于表 13-20 中的数据, 给出 Poisson-Poisson 分布参数的矩估计, 其中的次分布是普通的 (不是零截断的)Poisson 分布. 用选定的模型进行卡方拟合优度检验.

13.19 现有表 13-13 所示的数据, 这些数据显示了 23 589 份车险保单的结果. 第三列 "拟合模型" 给出了由 (最大似然估计确定参数) 负二项分布计算的期望损失数.

(a) 在 5%的显著性水平下进行卡方拟合优度检验.

(b) 求负二项参数 r 和 β 的最大似然估计. 可由已知数据直接计算而不必将似然函数最大化.

表 13-13 习题 13.19 的数据

损失次数 k	保单数 n_k	拟合模型
0	20 592	20 596.76
1	2 651	2 631.03
2	297	318.37
3	41	37.81
4	7	4.45
5	0	0.52
6	1	0.06
$\geqslant 7$	0	0.00

13.5 模 型 选 择

13.5.1 引言

我们已经介绍了几乎所有用于模型选择的工具. 在描述本书推荐的方法前, 必须介绍两个重要的概念. 首先是节俭(parsimony) 的原则, 节俭原则叙述为: 在没有足够充分的理由支持复杂模型时, 简单模型为更好的选择. 理由是, 虽然复杂模型和已知数据匹配得很好, 但并不能保证和观测值的总体匹配得很好. 例如, 任意给定 10 个有序数对 (x, y) 且 x 值互不相同, 则一定有一个不超过 9 次的多项式图形经过所有的 10 个点. 但如果上述点是随机样本, 总体的取值都在多项式的曲线上的可能性极小. 而且, 有可能存在一条直线不仅和样本点非常接近, 也和总体中的其他点非常接近. 这一点和大多数假设检验方法的思想是一致的, 即: 如果没有十分有力的证据, 就不要拒绝零假设 (而提出一个比零假设更复杂的对总体分布的描述).

第二个重要概念没有一个特定的名称, 它表述为: 如果尝试了足够多的模型,

即使所有模型都不足够好，总会有一个模型看上去是好的. 假设现在有 900 个模型可用，那么对于大多数的数据集，很可能会有一个模型拟合效果相当好，但是这并不会对我们了解总体的性质有所帮助.

综上所述，在选择模型的时候，必须明确两点：

(1) 只要可能，就尽量选择简单的模型；

(2) 限制备选模型的范围.

本节余下内容介绍的方法会对第一条的实施有所帮助，至于第二条，这就需要一些经验了 —— 特定的模型会在特定的情形下显得更加合适，只有丰富的经验才会提高建模者的判断力，从而缩减候选模型的名单.

本节内容将分别介绍两类模型选择标准. 第一类标准是基于建模者的主观判断，而第二类标准更为规范，使得大多数情况下所有人都会通过分析得出相同的结论，因为后者的结论不是由图形和表格的直观印象得来，而是通过对数据的定量分析得到的.

13.5.2 主观判断法

在根据个人的判断选择模型的过程中，必然至少要涉及以下三种观念之一. 但无论如何，分析员的经验都是至关重要的.

首先，可以根据本章介绍的各种图表工具 (或者基于图形的表格) 进行基本的判定①. 这使得进行真正的分析时可以集中在关系重大的方面. 例如，有时候对尾概率的拟合非常重要，而有时候众数的匹配更重要. 即便是评分法也可以运用一幅具有说服力的图形来支持选定的模型.

其次，决策也可能会受到在相似情景下成功采用某些特定的模型的影响，也有可能受到使用目的的影响. 比如，1941 年的美国保险委员会标准生命表 (Commissioners' Standard Ordinary Mortality Table, CSO) 在许多年龄段都采用了 Makeham 分布. 在计算能力有限的年代，这种分布使得对联合生存相关值的计算大大简化. 只要模型的拟合效果尚可接受，人们就会更加看中它简化计算的优势，而不采用拟合效果更优的其他模型. 类似地，如果某个分析师所在的公司或其他人一直使用 Pareto 分布作为某种责任保险 (liability insurance) 的损失模型，那么当采用其他分布时，可能就需要提出比通常情况更充足的证据才行.

再次，分布可能完全由具体情形决定. 例如，某口腔健康保险合约为投保人提供每年至多两次口腔检查，而投保个体每年会有两次独立的选择，决定是否进行检查，如果每次决定检查的概率是 q，则必然服从 $m=2$ 的二项分布.

① 除了已讨论的图表，还有很多其他的图或表格可以运用，例如 $q-q$ 图为理论模型与经验分布的有限期望函数或者平均剩余寿命函数的对比图.

最后必须注意, 以下介绍的通过严格的计算选择模型的方法, 得到的结果并不一定一致, 这时, 必然要通过主观判断来决定最终的选择.

13.5.3 评分法

某些分析师可能更喜欢通过自动过程来选择模型, 一种简单的方法是为每个模型打分, 得到最高分值的模型 "胜出". 以下一些打分的方法值得一试.

(1) Kolmogorov-Smirnov 检验统计量的最小值.

(2) Anderson-Darling 检验统计量的最小值.

(3) 卡方拟合优度检验统计量的最小值.

(4) 卡方拟合优度检验 p 值的最大值.

(5) 似然函数在最大似然估计点的取值的最大值.

除了卡方 p 值的方法外, 其余所有方法都在节俭原则上有缺陷. 首先, 考虑似然函数, 当比较指数模型和 Weibull 模型时, Weibull 模型的似然函数值必然不小于指数模型的似然函数值, 而且这两个值相等当且仅当 Weibull 参数 τ 的最大似然估计等于 1, 这种情况是很少见的. 因此 Weibull 分布总会胜过指数分布, 这显然违背了节俭原则. 对于其他 3 个检验统计量, 一般情况下并没有类似的关系. 不过似乎选择更为复杂的模型时, 拟合程度很可能更好. p 值方法能够不受模型复杂程度的干扰的唯一原因是: 随着模型复杂度的增加, 检验的自由度会减少, 进而较复杂的模型也可能有较小的 p 值. 对以上列出的前 2 个检验统计量并没有这种相对调整的措施.

而在运用似然函数值时, 又有两种操作方法. 一种是进行似然比检验, 另一种是在引入新的参数时有一定的惩罚. 只有当一个模型是另一个模型的特例时, 似然比检验在技术上才是可行的 (例如 Pareto 模型和一般化的 Pareto 模型). 通过使用显著水平 5%的检验, 这个概念可以变成一个算法. 从最好的单参数的模型 (有最大似然函数值的模型) 开始, 和两参数模型比较, 如果两参数模型的最大似然函数值增加了至少 1.92(将这个增量加倍就得到临界值的增量 3.84), 则采用两参数模型. 再考虑三参数模型, 如果和两参数模型比较, 仍然需要 1.92 的增量, 而如果第一步保留了单参数模型, 则三参数模型必须有 3.00 的增量才能被采用 (因为该检验的自由度是 2). 若要增加 3 个参数, 则需要 3.91 的增量, 4 个参数需要 4.74 的增量, 等等. 根据本章的思想, 这种法则也可以在模型之间没有包含关系的情形中使用, 但是可能无法认为这种情形下仍然在进行似然比检验.

除特殊情形外, 似然比检验也有其他假设检验相同的问题. 当样本容量加倍时, 对数似然函数值也会加倍, 使得参数个数更多的模型更容易被选择. 这将导致违背节俭原则的结果出现. 另一方面, 也可以认为, 如果掌握的数据相当多, 考虑更加复杂的模型也是理所当然的. 一种对上述立场的折中方法是 Schwarz Bayesian 准则

(SBC)[121], 这种方法在衡量模型时将对数似然函数值减去 $(r/2)\ln n$, 其中 r 是待估参数的个数, n 是样本容量.① 这样一来, 只有对数似然函数增加 $0.5\ln n$ 才能增加一个参数, 样本容量越大, 要求的似然函数增量就越大, 但这个要求的增量并非与样本容量成比例②.

例 13.13 (续例 13.12), *在指数分布和 Weibull 分布之间为数据选择模型.*

解 各个例题和习题都需要借助图表分析. 表 13-14 总结了各种模型选择数值的度量. 对于截断数据集 B, SBC 基于样本容量 19 进行计算; 而对于在 1 000 处删失的数据集, 共有 20 个观测值. 对于数据集 B 的以上两个形式, Weibull 模型都显示出了一定的优势, 不过没有很强的说服力. 特别地, 不管是似然比检验和 SBC 都没有体现出第二个参数的必要性. 而对于数据集 C, Weibull 模型有着明显的优势, 并且拟合效果相当好. □

表 13-14 例 13.13 的结果

	在 50 处截断的数据集 B		在 1 000 处删失的数据集 B	
比较标准	指数	Weibull	指数	Weibull
K-S*	0.134 0	0.088 7	0.099 1	0.099 1
A-D*	0.429 2	0.163 1	0.171 3	0.171 2
χ^2	1.403 4	0.361 5	0.595 1	0.594 7
p 值	0.843 6	0.948 1	0.897 6	0.742 8
对数似然函数	−146.063	−145.683	−113.647	−113.647
SBC	−147.535	−148.628	−115.145	−116.643
		数据集 C		
χ^2	61.913	0.369 8		
p 值	10^{-12}	0.946 4		
对数似然函数	−214.924	−202.077		
SBC	−217.350	−206.929		

* K-S 和 A-D 分别代表 Kolmogorov-Smirnov 检验统计量和 Anderson-Darling 检验统计量.

例 13.14 *在例 4.57 中使用了一种特别的方法说明 Poisson-ETNB 分布拟合得很好. 试用本章介绍的方法确定一个好的模型.*

解 由于数据集相当庞大, 因此需要模型和数据对应得很好. 计算结果见表 13-15.

由表 13-15 可见, 负二项分布拟合效果不佳. 而 Poisson- 逆高斯分布也只是好了一点点 ($p = 2.88\%$). Poisson- 逆高斯分布是 Poisson-ETNB 模型的一个特例 ($r = -0.5$). 因此, 可以用似然比检验来决定增加参数 r 是否恰当. 由于对数似然

① 在本书第一版中, 不仅 Schwarz 被拼错, 惩罚公式也是错误的, 本版纠正了这些错误.

② 也有其他一些基于已知信息的判定准则, 比如 Brockett [17] 第 3 节提出了 Akaike 信息准则 (Akaike information criterion). 在对这篇论文的讨论中 Carlin 提出了对 SBC 的支持观点.

表 13-15 例 13.14 的结果

索赔数	观测频数	用于拟合的分布		
		负二项	Poisson- 逆高斯	Poisson-ETNB
0	565 664	565 708.1	565 712.4	565 661.2
1	68 714	68 570.0	68 575.6	68 721.2
2	5 177	5 317.2	5 295.9	5 171.7
3	365	334.9	344.0	362.9
4	24	18.7	20.8	29.6
5	6	1.0	1.2	3.0
6+	0	0.0	0.1	0.4
参数		$\beta = 0.035\ 066\ 2$	$\lambda = 0.123\ 304$	λ=0.123 395
		$r = 3.577\ 84$	$\beta = 0.071\ 202\ 7$	$\beta = 0.233\ 862$
				$r = -0.846\ 872$
卡方		12.13	7.09	0.29
自由度		2	2	1
p 值		<1%	2.88%	58.9%
负对数似然函数		251 117	251 114	251 109
SBC		−251 130	−251 127	−251 129

函数增加了 5, 增量大于 1.92, 因此三参数模型的拟合效果有着明显的优势. 卡方检验的结果显示 Poisson-ETNB 已经给出了充分的拟合. 而另一方面, SBC 更倾向于选择 Poisson- 逆高斯分布, 它相对于三参数模型而言, 对尾概率的拟合效果更好, 因此看上去像是最佳选择. □

例 13.15 *这个例题选自 Douglas[29] 第 253 页. 某保险公司的年度记录显示了某一特定险种每天的索赔事故数, 如表 13-16 所示. 判断 Poisson 模型是否适合这些数据.*

表 13-16 例 13.15 的数据

索赔数/天	观测到的天数
0	47
1	97
2	109
3	62
4	25
5	16
6	4
7	3
8	2
9+	0

解 运用 Poisson 模型进行数据拟合. 由矩估计和最大似然方法都可以得到均值的

估计值

$$\hat{\lambda} = \frac{742}{365} = 2.032\,9.$$

卡方拟合优度检验的结果如表 13-17 所示. 无论何时, 这种表格所示的最后一组的期望数都是

$$E_{k+} = n\hat{p}_{k+} = n(1 - \hat{p}_0 - \cdots - \hat{p}_{k-1}).$$

最后三组被合并为一组, 来确保每一行的期望值至少为 1. 检验统计量为 9.93, 自由度为 6, 5%显著性水平的临界值为 12.59, p 值为 0.127 7. 由该项检验知 Poisson 分布是一个可接受的模型; 但是, 需要注意的是, 拟合效果在较大数值点显得很差, 因此采用该模型来解释会低估观测值, 因此是一个风险很大的选择. □

表 13-17 例 13.15 的卡方拟合优度检验

索赔数/天	观测值	期望值	卡方
0	47	47.8	0.01
1	97	97.2	0.00
2	109	98.8	1.06
3	62	66.9	0.36
4	25	34.0	2.39
5	16	13.8	0.34
6	4	4.7	0.10
7+	5	1.8	5.66
总计	365	365	9.93

例 13.16 表 12-13 的数据集选自 Beard et al.[12], 在例 12.58 中已经分析过. 找出一个能够充分描述这些数据的模型.

解 表 12-13 中有 4 种拟合模型的参数估计. 表 13-18 中有各种拟合的度量值. 只有零点调整的几何分布通过了拟合优度检验. 同时从 SBC 来看, 该分布显然也是最优的. 将该分布同几何分布进行似然比检验, 检验统计量为 2(171 479 − 171 133) = 692, 在一个自由度下, 显然是显著的. 例 12.58 中的定性的结论得到了确认. □

表 13-18 例 13.16 的结果

	Poisson	几何	ZM Poisson	ZM 几何
卡方	543.0	643.4	64.8	0.58
自由度	2	4	2	2
p 值	$< 1\%$	$< 1\%$	$< 1\%$	74.9%
对数似然	−171 373	−171 479	−171 160	−171 133
SBC	−171 379.5	−171 485.5	−171 173	−171 146

例 13.17 表 13-19 中的数据选自 Simon [122], 显示了 298 份合约的索赔数. 试给出一个合理的模型.

表 13-19 对 Simon 数据的拟合

索赔数/合约	合约数目	用于拟合的分布		
		Poisson	负二项	Polya-Aeppli
0	99	54.0	95.9	98.7
1	65	92.2	75.8	70.6
2	57	78.8	50.4	50.2
3	35	44.9	31.3	32.6
4	20	19.2	18.8	20.0
5	10	6.5	11.0	11.7
6	4	1.9	6.4	6.6
7	0	0.5	3.7	3.6
8	3	0.1	2.1	2.0
9	4	0.0	1.2	1.0
10	0	0.0	0.7	0.5
11	1	0.0	0.4	0.3
12+	0	0.0	0.5	0.3
参数		λ=1.708 05	β=1.159 07	λ=1.105 51
			r=1.473 64	β=0.545 039
卡方		72.64	4.06	2.84
自由度		4	5	5
p 值		< 1%	54.05%	72.39%
对数似然		−577.0	−528.8	−528.5
SBC		−579.8	−534.5	−534.2

解 分别用 Poisson、负二项和 Polya-Aeppli 分布对数据进行拟合, 发现 Polya-Aeppli 和负二项都是表面上可接受的. 卡方统计量 p 值以及对数似然函数都表明 Polya-Aeppli 模型略优于负二项模型. 而 SBC 确认了这两个模型都优于 Poisson 分布. 要作出最终的决定, 需要视模型使用者对负二项和 Polya-Aeppli 两个模型的熟悉程度、使用经验以及计算是否简便而定. □

例 13.18 考虑如表 13-20 所示的摘自 Bühlmann[19] 的瑞士车险数据, 试给出一个合理的模型.

解 表 13-20 考虑了 3 个模型. 其中, Poisson 分布的拟合效果非常差, 它的尾部和实际经验相比显得实在太薄了. 负二项分布看上去改进了许多, 但是由于卡方统计量的 p 值太小而不能接受. 如此大的样本容量自然要求更好的拟合效果. 用 Poisson- 逆高斯分布进行拟合的效果近乎完美 (p 值很大). 注意到 Poisson- 逆高斯分布和负二项分布一样有 2 个参数. SBC 也显示出 Poisson- 逆高斯分布较好. 从这个例子可以看出, Poisson- 逆高斯分布的右端尾部要比负二项分布厚得多. □

例 13.19 Bevan[15] 于 1963 年全面研究了医疗保险的索赔, 其中包含男性 (955 例赔付) 和女性 (1 291 例赔付) 的索赔, 数据见表 13-21, 其中免赔额是 25. 可以用

一个共同的模型来拟合男性和女性数据吗?

表 13-20 对 Bühlmann 数据的拟合

事故数	观测频数	用于拟合的分布		
		Poisson	负二项	P.-i.G.[a]
0	103 704	102 629.6	103 723.6	103 710.0
1	14 075	15 922.0	13 989.9	14 054.7
2	1 766	1 235.1	1 857.1	1 784.9
3	255	63.9	245.2	254.5
4	45	2.5	32.3	40.4
5	6	0.1	4.2	6.9
6	2	0.0	0.6	1.3
7+	0	0.0	0.1	0.3
参数		$\lambda = 0.155\ 140$	$\beta = 0.150\ 232$	λ=0.144 667
			$r = 1.032\ 67$	β=0.310 536
卡方		1 332.3	12.12	0.78
自由度		2	2	3
p 值		<1%	<1%	85.5%
对数似然		−55 108.5	−54 615.3	−54 609.8
SBC		−55 114.3	−54 627.0	−54 621.5

a. P.-i.G. 代表 Poisson- 逆高斯.

表 13-21 例 13.19 的医疗保险损失额

损失额	男性	女性
25~50	184	199
50~100	270	310
100~200	160	262
200~300	88	163
300~400	63	103
400~500	47	69
500~1 000	61	124
1 000~2 000	35	40
2 000~3 000	18	12
3 000~4 000	13	4
4 000~5 000	2	1
5 000~6 667	5	2
6 667~7 500	3	1
7 500~10 000	6	1

解 当运用合并的数据集进行建模时, 对数正态分布是最好的两参数模型, 其负对数似然函数 (NLL) 为 4 580.20, 这比单参数的逆指数模型要小 19.09, 比三参数

的 Burr 模型要大 0.13. 由于这些模型没有某个模型是另一个模型的特例, 因此不能利用似然比检验 (LRT). 但是可以明显地看到, 用 1.92 作为差值标准, 对数正态模型更优. SBC 要求增加参数时其改进至少为 $0.5\ln(2\ 264) = 3.86$, 此时对数正态模型还是更优的选择. 参数是 $\mu = 4.523\ 7, \sigma = 1.495\ 0$. 如果使用男性数据和女性数据分别建立对数正态模型 (男性模型 $\mu = 3.968\ 6, \sigma = 1.843\ 2$, 女性模型 $\mu = 4.771\ 3, \sigma = 1.284\ 8$), NLL 分别为 1 977.25 和 2 583.82, 总计 4 561.07. 这比共同的对数正态模型增加了 19.13, 用 LRT 判断 (需要增加 3.00) 和用 SBC 判断 (需要增加 7.72) 都很显著. 有的时候可能会在两个模型中使用相同的非尺度参数, 如果使用公共的 σ 值, 则 NLL 是 4 579.77, 显著地差于单独的模型. □

例 13.20 Longley-Cook[86] 于 1958 年调查了非寿险精算师的从业状况, 列表给出了非寿险公司在 1949 年 (共 55 位精算师) 以及 1957 年 (共 78 位精算师) 聘用的美国非寿险精算学会 (Casualty Actuarial Society, CAS) 会员的数目. 基于表 13-22 的数据, 对各个 (至少聘用了一位精算师的) 公司的精算师数目建立模型, 并判断其分布在 8 年的时间中是否发生了变化.

解 由于取值为零的情况不可能出现, 因此只考虑在零点截断的分布. 不管在何种情况下 (只考虑 1949 年的数据, 或只考虑 1957 年的数据或同时考虑两组数据), 在零点截断的对数分布以及在零点截断的 (推广的) 负二项分布的拟合优度检验值都是可以接受的. NLL 的改进量是 0.52, 0.02 和 0.94. 也可运用 LRT 方法 (除了零点截断的对数分布是零点截断的负二项分布在 $r \to 0$ 时的极限情形), 可发现改进都不显著, 用 SBC 也可得到相同的结论. 参数估计 (β 是唯一的参数) 的结果分别是 2.022 7, 2.811 4 和 2.447 9. 合并数据集的 NLL 是 74.35, 分别建立的模型的和是 74.15, 改进量是 0.20, 也是不显著的 (自由度 1). 即使均值的估计值从 $2.022\ 7/\ln(3.022\ 7) = 1.828\ 6$ 增加为 $2.811\ 4/\ln(3.811\ 4) = 2.101\ 2$, 但并没有足够的数据来得出真实的均值确实增加了的结论. □

表 13-22 例 13.20 中每个公司的精算师人数

精算师人数	公司数 —1949	公司数 —1957
1	17	23
2	7	7
3~4	3	3
5~9	2	3
10+	0	1

习题

13.20* 现有 1 000 份保单的事故数记录如表 13-23 所示. 在没有进行任何正规的假设检验前, 判断以下各模型的适用性: 二项, Poisson, 负二项, 正态以及 gamma.

表 13-23 习题 13.20 的数据

事故数	保单数
0	100
1	267
2	311
3	208
4	87
5	23
6	4
总计	1 000

13.21 判断例 13.1 的变形的 gamma 分布在 3 个数据集上是否优于指数分布或 Weibull 分布.

13.22* 习题 13.11 数据的 Poisson 分布最大似然估计为 $\hat{\lambda} = 0.60$, 负二项分布最大似然估计为 $\hat{r} = 2.9$ 和 $\hat{\beta} = 0.21$. 用似然比检验法在这两个模型中进行选择.

13.23* 现采用 5 个模型对容量为 100 的样本进行拟合, 结果见表 13-24. 用 Schwarz Bayesian 准则选择最佳的模型.

表 13-24 习题 13.23 的结果

模　型	参数个数	负对数似然函数
广义 Pareto	3	219.1
Burr	3	219.2
Pareto	2	221.2
对数正态	2	221.4
逆指数	1	224.3

13.24 续习题 12.38. 同时用似然比检验 (显著性水平 5%) 和 Schwarz Bayesian 准则判断 Sylvia 的观点是否正确.

13.25 利用习题 12.96 和 13.15 的结果, 用卡方拟合优度检验, 似然比检验以及 Schwarz Bayesian 准则找出 $(a,b,0)$ 类中的最佳模型.

13.26 利用习题 12.98 和 13.16 的结果, 用卡方拟合优度检验, 似然比检验以及 Schwarz Bayesian 准则找出 $(a,b,0)$ 类中的最佳模型.

13.27 利用习题 12.99 和 13.17 的结果, 用卡方拟合优度检验, 似然比检验以及 Schwarz Bayesian 准则找出 $(a,b,0)$ 类中的最佳模型.

13.28 表 13-25 给出了每次汽车事故报告中医疗保险的索赔次数.

(a) 绘出类似图 4-8 的散点图. 由图判断是否为 $(a,b,0)$ 类成员? 若是, 为哪个分布?

(b) 对 $(a,b,0)$ 类的每一个分布, 求其参数的最大似然估计.

(c) 根据卡方拟合优度检验, 似然比检验以及 Schwarz Bayesian 准则, $(a,b,0)$ 类中的哪个分布拟合效果最好? 该模型是可接受的吗?

表 13-25 习题 13.28 的数据

医疗保险索赔	事故数
0	529
1	146
2	169
3	137
4	99
5	87
6	41
7	25
8+	0

13.29 假设对习题 12.96、12.98、12.99 和 13.28 中涉及的 4 个数据集, 已经求出了 (*a*,*b*,0) 类分布中最佳的模型. 试对每个数据集, 求以下模型的最大似然估计: 零点调整的 Poisson, 几何, 对数以及负二项分布. 用卡方拟合优度检验以及似然比检验判断所考虑的 8 个模型中的最佳模型, 并说明选定的模型是否可以接受.

13.30 到目前为此我们还没有介绍的一个频率模型是**zeta 分布**, 该分布在零点截断, 概率函数为 $p_k^T = k^{-(\rho+1)}/\zeta(\rho+1), k = 1,2,\cdots,\rho > 0$ 其中的分母表示 zeta 函数, 由 $\zeta(\rho+1) = \sum_{k=1}^{\infty} k^{-(\rho+1)}$ 进行数值计算得到. 也可以用通常的办法构造零点调整的 zeta 分布, 关于该分布的更多信息请参阅 Luong and Doray[88].

(a) 基于例 12.58 中的数据, 计算零点调整 zeta 分布参数的最大似然估计.

(b) 零点调整的 zeta 分布是可接受的模型吗?

13.31 在习题 13.29 中, 已经求出了 (*a*,*b*,0) 类和 (*a*,*b*,1) 类分布函数中基于习题 12.96, 12.98, 12.99 和 13.28 的数据的最佳模型. 现用 Poisson-Poisson, Polya-Aeppli, Poisson- 逆高斯以及 Poisson-ETNB 分布拟合这些数据, 并判断这些模型是否能代替习题 13.29 中选定的模型. 新得到的最佳模型是可接受的吗?

13.32 本题用到的 5 个数据集都来自于 Lemaire[82]. 计算每个数据集的前三阶矩, 然后用 4.6.8 节的思想在复合 Poisson 族模型 [Poission, 几何, 负二项, Poisson- 二项 ($m = 2$ 和 $m = 3$), Polya-Aeppli, Neyman A 型, Poisson- 逆高斯以及 Poisson-ETNB] 中进行选择. 从选定的模型 (如果有的话) 以及 (*a*,*b*,0) 类和 (*a*,*b*,1) 类分布中选择最佳模型.

(a) 表 13-26 的数据为比利时机动车第三者责任险的索赔数目.

(b) 表 13-27 的数据为 1875 年到 1894 年间普鲁士军队中因马蹬踢而死亡的人数. 记录了在给定年份里 1 个军团 (共有 10 个军团) 中的死亡人数, 因此共有 200 个观测值. 在介绍 Poisson 分布的最初想法时引用过这个数据集. 在运用我们的模型的时候, 还需要对数据作什么样的进一步假设?

(c) 表 13-28 的数据为 1500 年到 1931 年间各年主要的国际战争数.

(d) 表 13-29 的数据为 1947 年到 1960 年间的世界棒球大赛中每半局的得分数.

(e) 表 13-30 的数据为 1966—1967 赛季中全美曲棍球联赛每支队伍每场比赛的得分.

表 13-26 习题 13.32(a) 的数据

索赔数	保单数
0	96 978
1	9 240
2	704
3	43
4	9
5+	0

表 13-27 习题 13.32(b) 的数据

死亡人数	军团数
0	109
1	65
2	22
3	3
4	1
5+	0

表 13-28 习题 13.32(c) 的数据

战争数	年份数
0	223
1	142
2	48
3	15
4	4
5+	0

表 13-29 习题 13.32(d) 的数据

得分数	半局数
0	1 023
1	222
2	87
3	32
4	18
5	11
6	6
7+	3

表 13-30 习题 13.32(e) 的数据

得分数	比赛场数
0	29
1	71
2	82
3	89
4	65
5	45
6	24
7	7
8	4
9	1
10+	3

13.33 验证例 4.64 给出的估计值是最大似然估计. (由于只保留了两位小数, 似然函数在附近点的取值可能略小.) 在例 12.56 中, 又使用了负二项分布对数据进行拟合. 以上两个模型, 哪一个更可取?

第14章 实　例

14.1 引　言

本章通过 5 个实例来说明迄今为止所讨论的许多概念. 第一个例子是死亡时间模型. 第二个例子是医疗事故从发生到报告的时间模型. 第三个例子是责任险赔付金额模型, 这个模型也是连续的, 但一般具有递减的失效率 (典型的是赔付额变量). 而另一方面, 某事件的发生时间变量的失效率一般为递增的. 对最后两个例子还讨论了第 6 章的总损失模型以及混合模型.

14.2 死亡时间

14.2.1 数据

SOA 网站www.soa.org公布了不同的死亡表. 一般死亡表将给出整数死亡年龄的生存函数. 表 14-1 为 1900 年的女性死亡率的部分数据. 图 14-1 将表 14-1 的点直接连线得到.

表 14-1　1900 年女性死亡率

x	$S(x)$	x	$S(x)$	x	$S(x)$
0	1.000	35	0.681	75	0.233
1	0.880	40	0.650	80	0.140
5	0.814	45	0.617	85	0.062
10	0.796	50	0.580	90	0.020
15	0.783	55	0.534	95	0.003
20	0.766	60	0.478	100	0.000
25	0.739	65	0.410		
30	0.711	70	0.328		

如果假设生存函数就是这样的一条连接各个点的直线, 则可以计算平均剩余生命函数. 由 (3.5) 式知, 可以用给定年龄的曲线下方的面积除以该年龄的生存函数计算. 图 14-2 为平均剩余生命函数的图形. 出生后不久的微量上升说明 1900 年婴儿的死亡率较高, 出生一年后的生存个体的期望剩余寿命会多 5 年. 随后的平均剩余寿命平稳递减, 这符合我们预期的衰老效应.

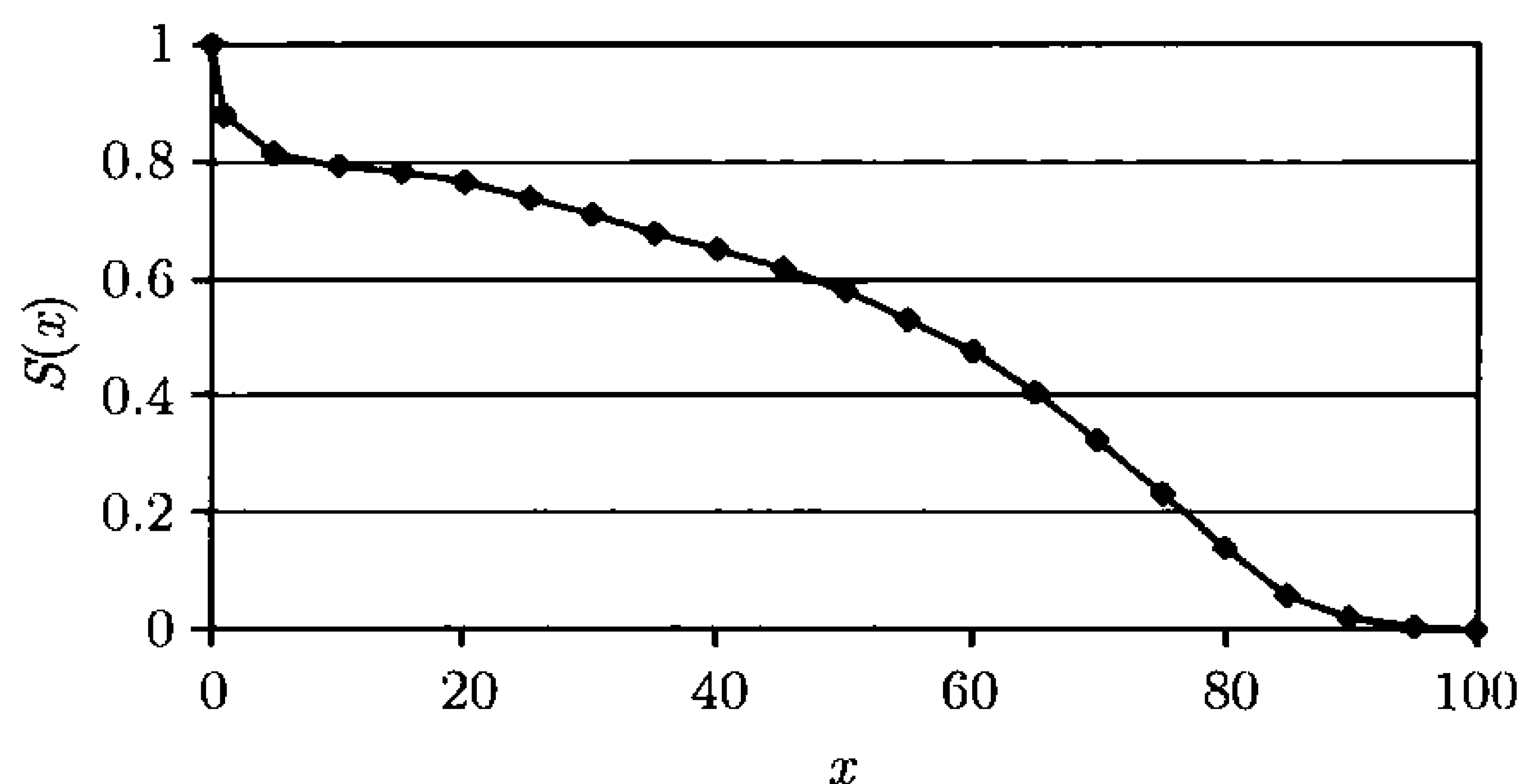


图 14-1 生存函数 (SOA 数据)

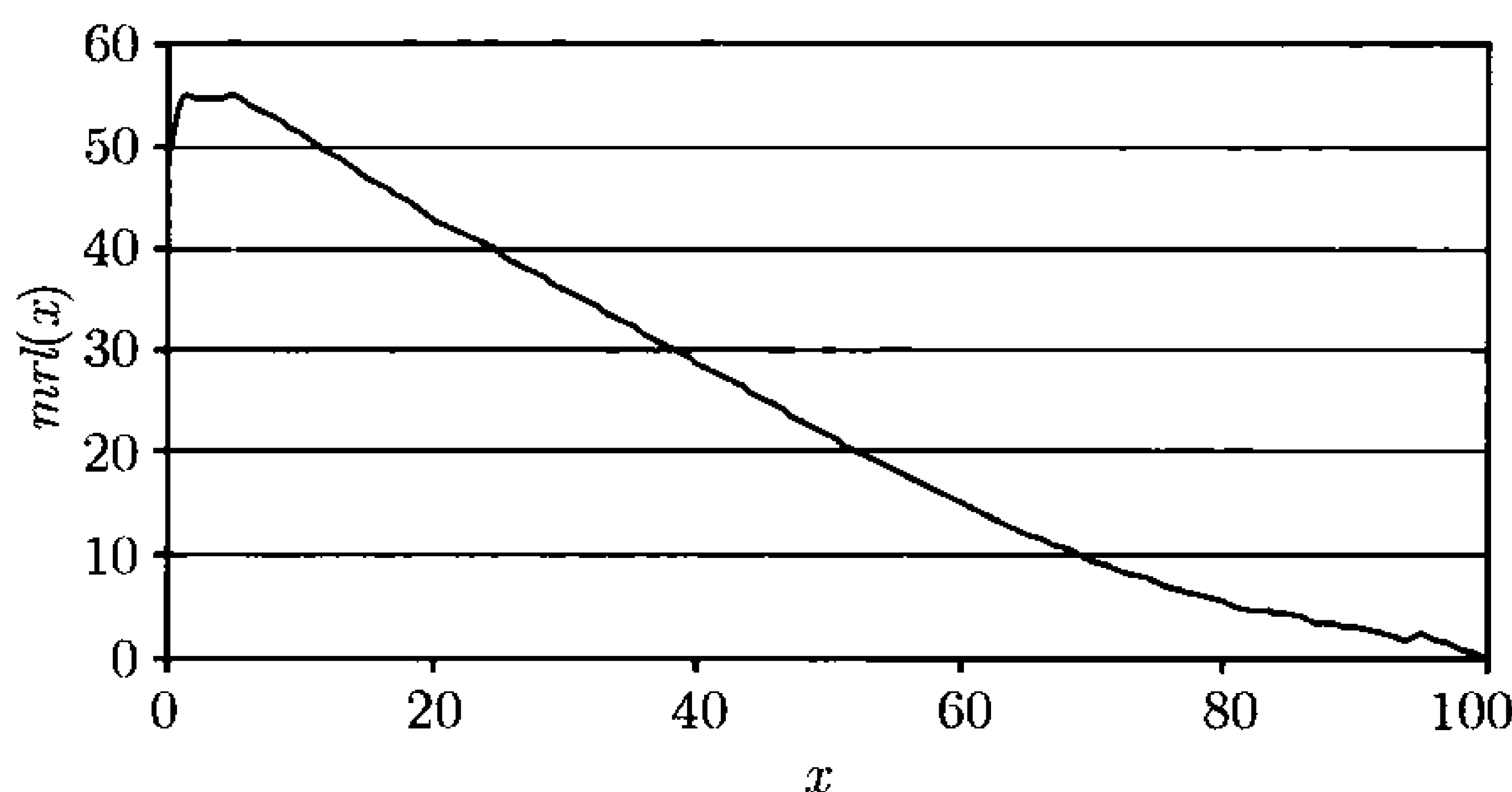


图 14-2 平均剩余生命函数 (SOA 数据)

14.2.2 基本计算

在人寿保险中, 人们并不特别关心例如免赔、限额和共保等条款. 这时, 将考虑如下两个问题.

(1) 当前 65 岁的个体, 在今后的生存期间每年年初受益 1 000 元, 并以 6%的年利率计算这些现金流的期望现值.

(2) 当前 20 岁的个体, 若在死亡时刻给付 1 000 元, 以年利率 6%计算该现金流的期望现值.

对于第一个问题, 现值随机变量可以表示为 $Y = 1\ 000(Y_0 + \cdots + Y_{34})$, 其中 Y_j 表示生存到 65+j 岁的条件下受益 1 个货币单位的现值. 则有

$$Y_j = \begin{cases} 1.06^{-j}, & \text{概率为 } \dfrac{S(65+j)}{S(65)}, \\ 0, & \text{概率为 } 1 - \dfrac{S(65+j)}{S(65)}. \end{cases}$$

结果为

$$\mathrm{E}(Y) = 1\ 000 \sum_{j=0}^{34} \frac{1.06^{-j} S(65+j)}{0.410}$$

$$= 8\ 408.07,$$

其中生存函数的中间值由线性插值计算.

对于第二个问题, 令 $Z = 1\ 000(1.06^{-T})$ 表示现值随机变量, T 表示 20 岁个体的未来生存时间, 以年为单位. 计算公式为

$$\mathrm{E}(Z) = 1\ 000 \int_0^{80} \frac{1.06^{-t} f(20+t)}{S(20)} \mathrm{d}t.$$

其中生存函数在整数年龄之间的值由线性插值得到, 密度函数为曲线的斜率. 即如果 x 是 5 的倍数, 则

$$f(t) = \frac{S(x) - S(x+5)}{5}, \quad x < t < x+5.$$

将积分区域划分成 16 段, 有

$$\begin{aligned}\mathrm{E}(Z) &= \frac{1\ 000}{0.766} \sum_{j=0}^{15} \frac{S(20+5j) - S(25+5j)}{5} \int_{5j}^{5+5j} 1.06^{-t} \mathrm{d}t \\ &= \frac{200}{0.766} \sum_{j=0}^{15} [S(20+5j) - S(25+5j)] \frac{1.06^{-5j} - 1.06^{-5-5j}}{\ln\ 1.06} \\ &= 155.10.\end{aligned}$$

这种情况通常不使用参数模型, 但这里考虑损失率函数为 $h(x) = A + Bc^x$ 的 Makeham 分布. 则

$$S(x) = \exp\left[-Ax - \frac{B(c^x - 1)}{\ln\ c}\right].$$

因为没有给出样本规模, 故无法用最大似然估计. 因为这个模型不可能对 20 岁以下的个体有效用, 所以只有超过这个年龄的信息才可使用. 假设生存函数服从表 14-1 的个体最初有 1 000 人, 则 30~35 岁的似然函数为 $30 \ln\{[S(30) - S(35)]/S(20)\}$, 其中的生存函数由 Makeham 分布计算. 通过表中的数据计算样本量为 1 000(0.711−0.681)[①], 则似然函数最大的参数值为 $\hat{A}$=0.006 698, $\hat{B}$=0.000 079 76, $\hat{C}$=1.095 63. 在图 14-3 中, 点表示数据, 曲线为 Makeham 生存函数 (两者都是在活过 20 岁的条件下). 这个拟合的效果太好了, 或许说明了成人死亡率表已经用 Makeham 分布光滑处理过.

---

① 除了没有样本规模的信息外, 表 14-1 的值本身可能也不是随机观测的, 可能已经用第 15 章讨论的方法进行了光滑处理.

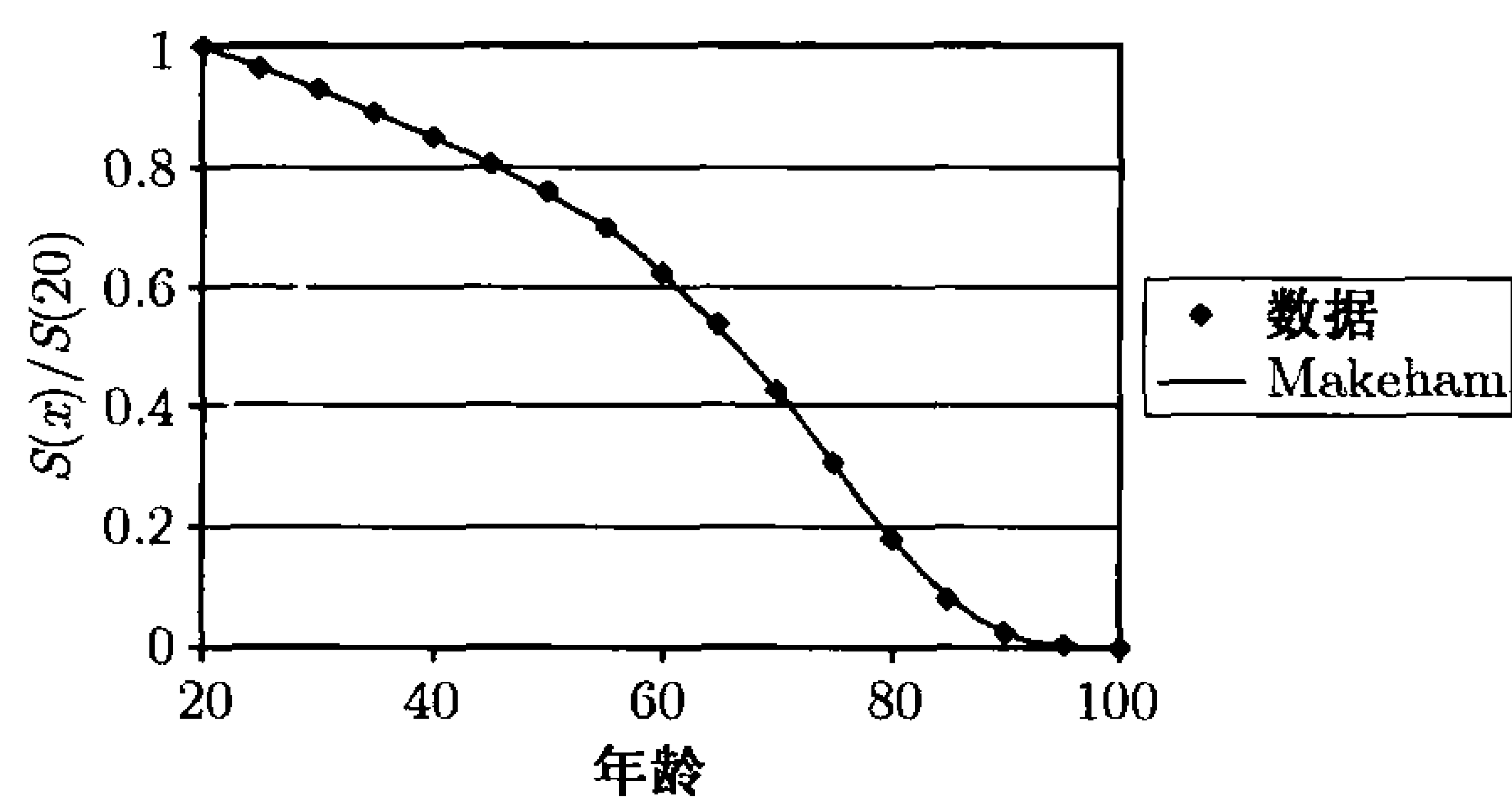


图 14-3 数据和 Makeham 模型的比较

还可以对其他险种进行类似的计算. 对于前面提出的年金, 不需要插值是因为 Makeham 分布给出了每个整数年龄的生存函数值. 第一个问题的现值为 8 405.24. 对于第二个寿险问题很难得到积分的解析表达式, 若在整数之间采用线性插值可以得到现值 154.90. 不难理解这些结果与之前公式的结果是一致的.

习题

14.1 年龄从 5 岁到 100 岁的平均剩余生命函数是近似线性的. 因为很少有 5 岁以下个体的寿险产品, 所以很自然地将曲线直接连线到零点, 进而合理地近似为 $e(x) = 60 - 0.6x$. 然后利用其确定死亡时间的密度函数和生存函数, 并且用这个函数计算前面的两个问题.

14.3 从事故发生到报告的时间

任何一个保险合同都是对特定的事件 (如死亡、伤残、火灾等) 提供赔付的. 这里有 3 个关键的时点. 第一个是事故发生的时间, 第二个是事故报告给保险公司的时间, 第三个是索赔支付的时间. 这些时间数据很重要, 因为它关系到索赔支付之前由保费产生的利息收入, 并提供了一种估算未报告索赔的机制. 这部分的例子将讨论从事故发生到报告的时间间隔, 主要是基于 Accomando and Weissner 的论文 [4] 来讨论的.

14.3.1 问题和数据

本例考虑某个特定年份的医疗事故索赔, 数据记录了从研究开始的某年年初起的 168 个月中的 463 起事故报告. 表 14-2 按每 6 个月 1 个区间给出了事件发生到索赔的时间间隔的分布. 平均剩余生命函数如图 14-4[①].

我们的目标是选择一个模型来拟合这些观测值, 并估计该年发生的总索赔额. 平均剩余生命函数的图像显示其是递减的, 并且比指数分布的尾部略轻一些.

① 因为数据右截尾, 故在计算平均剩余生命时丢掉了一些数据. 但从数据中不能明显地看出有什么影响. 这个图只是一个指导, 而最终选择的模型应该既很好地拟合数据也符合分析者的经验和判断.

Weibull 分布模型的尾部符合这样的特点, 可以用在这里.

表 14-2 医疗事故报告延迟

延迟的月份数	索赔数	延迟的月份数	索赔数
0~6	4	84~90	11
6~12	6	90~96	9
12~18	8	96~102	7
18~24	38	102~108	13
24~30	45	108~114	5
30~36	36	114~120	2
36~42	62	120~126	7
42~48	33	126~132	17
48~54	29	132~138	5
54~60	24	138~144	8
60~66	22	144~150	2
66~72	24	150~156	6
72~78	21	156~162	2
78~84	17	162~168	0

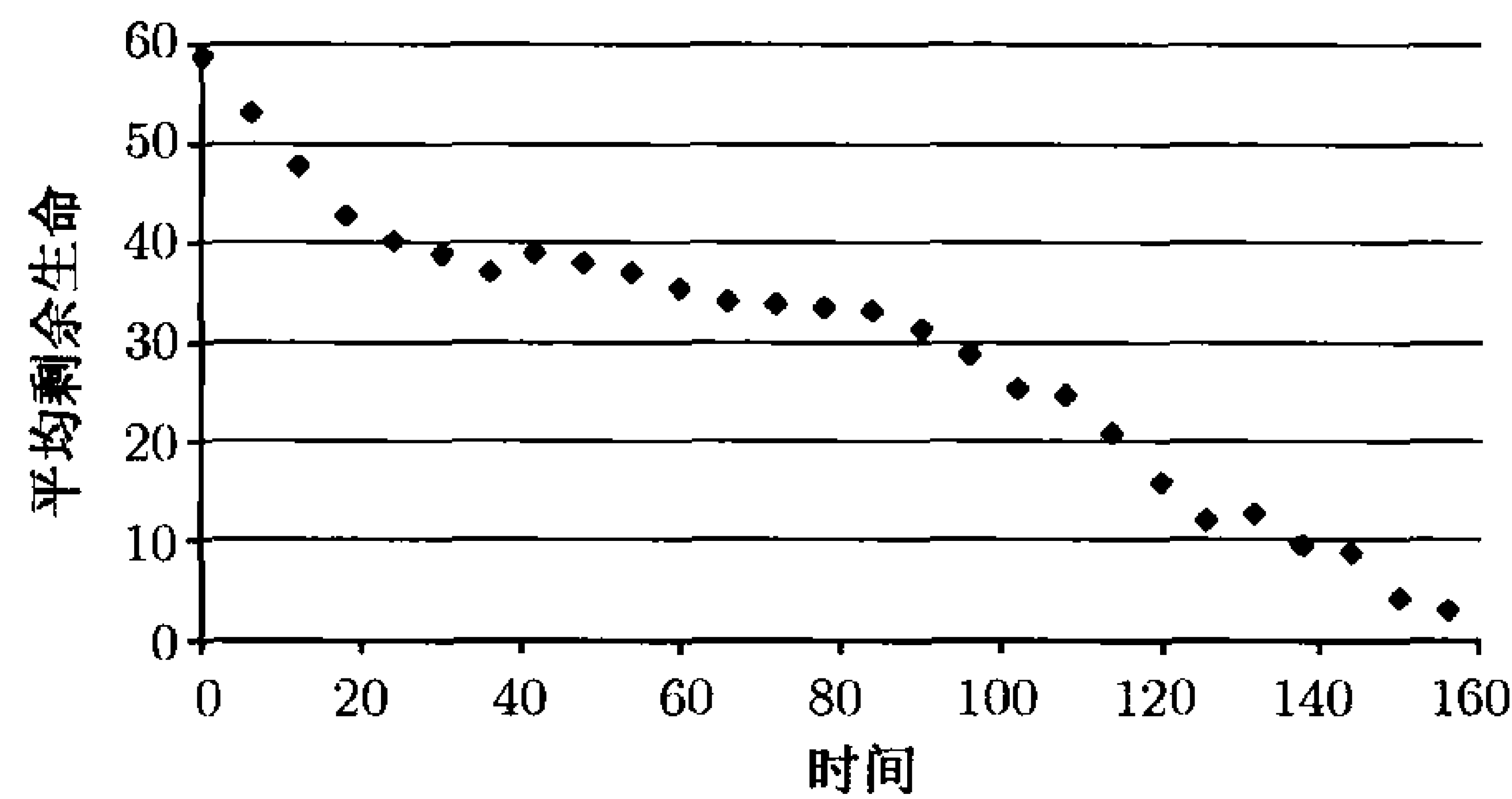


图 14-4 延迟报告数据的平均剩余生命函数

14.3.2 分析

由最大似然估计 Weibull 分布的参数为 $\hat{\tau}$=1.712 68, $\hat{\theta}$=67.300 2. 根据 Weibull 分布, 索赔报告时间不超过 168 的概率为

$$F(168) = 1 - \mathrm{e}^{-(168/\theta)^{\tau}}.$$

如果 N 是未知的总索赔次数, 则在 168 之前的观测数为二项分布, 因此在期望值的基础上得到, 时刻 168 之前报告的索赔期望次数为 $N[1 - \mathrm{e}^{-(168/\theta)^{\tau}}]$.

令这个期望值等于 463, 则 N 为

$$N = \frac{463}{1 - \mathrm{e}^{-(168/\theta)^{\tau}}}.$$

代入参数估计得到 466.88. 因此, 预期 14 年后会有 4 个以上的索赔报告.

由 Delta 模型 (定理 12.17) 可以构造 95%置信区间, 即 466.88 ± 2.90, 这说明索赔报告次数可能介于 1 个与 7 个之间.

14.4 赔 付 额

某再保公司要求咨询精算师帮助其确定各种保险责任的预期费用和风险 (用变异系数度量). 为此精算师收集了 200 个索赔的损失量, 再保险人还估计 (设这个估计是完全可信的) 每年约有 21 次损失, 损失次数服从 Poisson 分布. 我们所关心的是全额赔付情形, 250 000 以上最多为 100 万, 500 000 以上最多为 200 万. 这里所说的 "y 以上最多为 z" 的解释如定理 5.13 中的定义 $d = y$, $u = z + y$.

14.4.1 数据

表 14-3 列出了 178 个损失额不超过 200 000 的索赔 (1 000 美元为单位). 另外还有 22 个损失超过 200 的索赔如下:

206	219	230	235	241	272	283	286	312	319	385
427	434	555	562	584	700	711	869	980	999	1506

表 14-3 不超过 200 的损失 (1 000 美元为单位)

损失区间 (千元)	损失数	损失区间 (千元)	损失数
1~5	3	41~50	19
6~10	12	51~75	28
11~15	14	76~100	21
16~20	9	101~125	15
21~25	7	126~150	10
26~30	7	151~200	15
31~40	18		

表中 178 个损失的总和为 11 398, 平方和为 1 143 164.

为了帮助理解这些数据, 做直方图 14-5. 注意图中的高为每个区间的观测数除以样本个数 (200) 然后再除以区间宽度. 因此, 第一个柱的高度为 $3/[200(5)] = 0.003$.

从直方图看出, 潜在的分布有一个非零的众数点. 为了检查尾部的情况, 计算了一些点的经验平均剩余生命函数, 见表 14-4, 函数近似为常数, 因此指数模型比较合适.

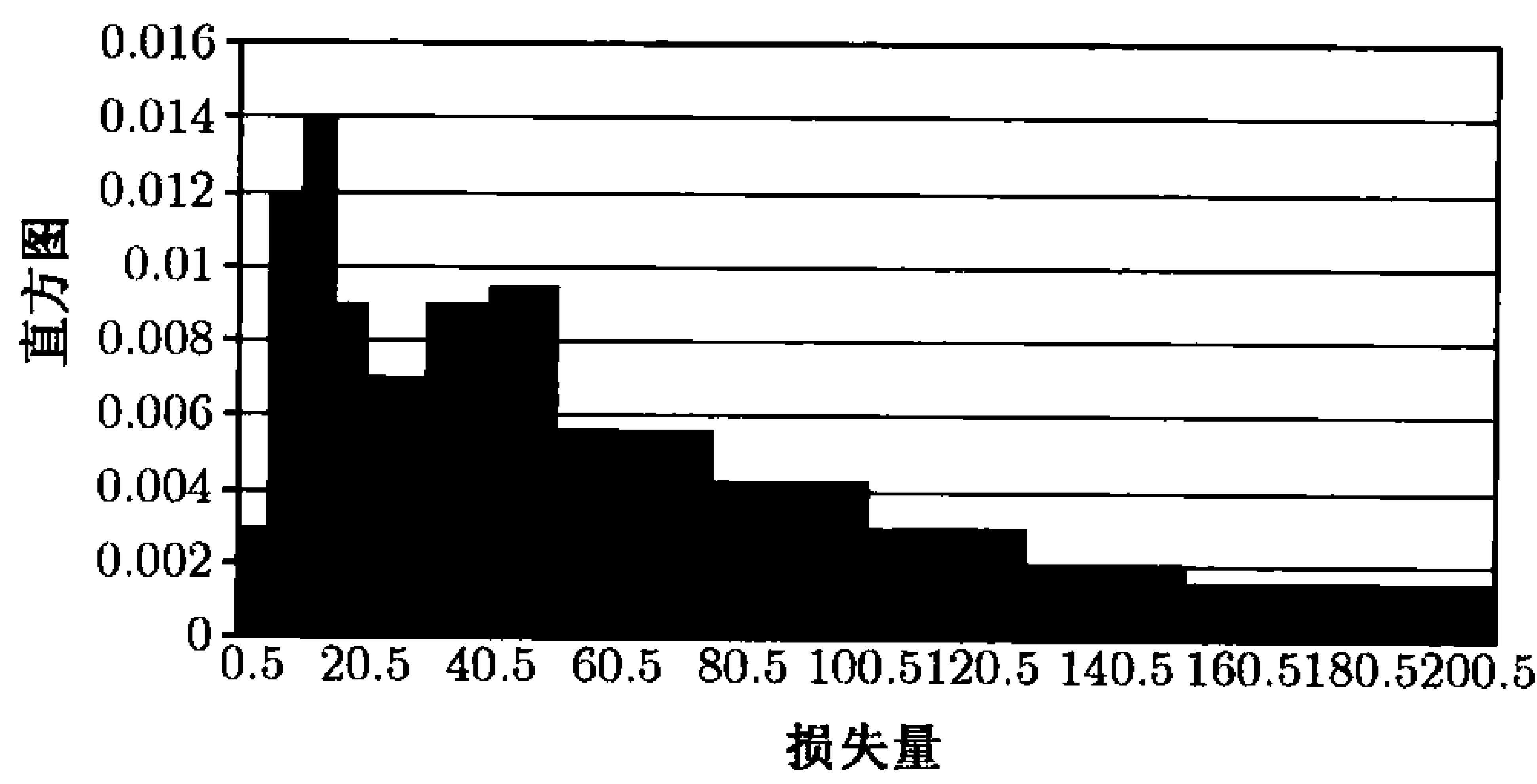


图 14-5 损失量的直方图

表 14-4 损失量超过 200 的平均剩余生命函数 (1 000 美元为单位)

损失	平均剩余生命
200	314
300	367
400	357
500	330
600	361
700	313
800	289
900	262

14.4.2 第一个模型

选择一个分两段的模型, 200(1 000 美元为单位) 之前为经验模型, 200 之后为指数模型. 至少有两种方法确定指数模型. 一种方法是限制参数, 使分布在超过 200 处的概率为 11%(22/200). 另一种方法是不考虑 11%的要求直接估计指数模型, 然后乘上密度函数使其超过 200 的区域为 0.11. 后者得到的参数估计为 $\theta = 314$. 位于 200 以下的部分, 经验分布在每个观测点的概率为 1/200. 则指数分布的密度函数 ($x > 200$) 为

$$f(x) = 0.000\ 662\ 344\mathrm{e}^{-x/314}.$$

对于损失量的所有区域, k 阶矩为 (样本中的 200 个损失由小到大排序)

$$\mathrm{E}(X^k) = \frac{1}{200}\sum_{j=1}^{178} x_j^k + \int_{200}^{\infty} x^k f(x)\mathrm{d}x.$$

则

$$\mathrm{E}(X)=\frac{11\ 398}{200}+0.000\ 662\ 344[314(200)+314^2]\mathrm{e}^{-200/314}=113.53,$$

$$\begin{aligned}\mathrm{E}(X^2)=&\frac{1\ 143\ 164}{200}\\&+0.000\ 662\ 344[314(200)^2+2(314)^2(200)+2(314)^3]\mathrm{e}^{-200/314}\\=&45\ 622.93.\end{aligned}$$

方差为 $45\ 622.93-113.53^2=32\ 733.87$, 变异系数为 1.59. 但这只是一次损失的情况. 年损失服从复合 Poisson 分布, 均值为

$$\mathrm{E}(S)=\mathrm{E}(N)\mathrm{E}(X)=21(113.53)=2\ 384.13,$$

方差为

$$\begin{aligned}\mathrm{Var}(S)&=\mathrm{E}(N)\mathrm{Var}(X)+\mathrm{Var}(N)\mathrm{E}(X)^2\\&=21(32\ 733.87)+21(113.53)^2=958\ 081.53,\end{aligned}$$

变异系数为 0.41.

对于其他的赔付, 需要给出前两阶限额矩 (limited expected moments) 的一般公式. 当 $u>200$, 有

$$\begin{aligned}\mathrm{E}(X\wedge u)&=56.99+\int_{200}^{u}xf(x)\mathrm{d}x+\int_{u}^{\infty}uf(x)\mathrm{d}x\\&=56.99+c\int_{200}^{u}x\mathrm{e}^{-x/314}\mathrm{d}x+c\int_{u}^{\infty}u\mathrm{e}^{-x/314}\mathrm{d}x\\&=56.99+c\left(-314x\mathrm{e}^{-x/314}-314^2\mathrm{e}^{-x/314}\right)\Big|_{200}^{u}\\&\quad+-cu314\mathrm{e}^{-x/314}\Big|_{u}^{\infty}\\&=56.99+c\left(161\ 396\mathrm{e}^{-200/314}-314^2\mathrm{e}^{-u/314}\right),\end{aligned}$$

其中 $c=0.000\ 662\ 344$, 类似地

$$\begin{aligned}\mathrm{E}[(X\wedge u)^2]&=5\ 715.82+c\int_{200}^{u}x^2\mathrm{e}^{-x/314}\mathrm{d}x+c\int_{u}^{\infty}u^2\mathrm{e}^{-x/314}\mathrm{d}x\\&=5\ 715.82+c[-314x^2-314^2(2x)-314^3(2)]\mathrm{e}^{-x/314}\Big|_{200}^{u}\\&\quad-cu^2314\mathrm{e}^{-x/314}\Big|_{u}^{\infty}\\&=5\ 715.82+c\left[113\ 916\ 688\mathrm{e}^{-200/314}-(197\ 192u+61\ 918\ 288)\mathrm{e}^{-u/314}\right].\end{aligned}$$

表 14-5 给出了完成所有工作需要的计算.

表 14-5 限额矩计算

u	$\mathrm{E}(X \wedge u)$	$\mathrm{E}[(X \wedge u)^2]$
250	84.07	12 397.08
500	100.24	23 993.47
1 250	112.31	41 809.37
2 500	113.51	45 494.83

若要求每次损失超过 250 限额为 1 000 的前两阶矩为

$$\begin{aligned}\text{均值} &= 112.31 - 84.07 = 28.24,\\ \text{二阶矩} &= 41\ 809.37 - 12\ 397.08 - 2(250)(28.24) = 15\ 292.29,\\ \text{方差} &= 15\ 292.29 - 28.24^2 = 14\ 494.79,\\ \text{协方差} &= \frac{\sqrt{14\ 494.79}}{28.24} = 4.26.\end{aligned}$$

如我们预期的一样, 保单限额的降低将减少方差, 同时由变异系数度量的风险将有大幅上升. 全年的均值为 593.04, 方差为 321 138.09, 变异系数为 0.96.

若要求每次损失超过 500 且限额为 2 000 的前两阶矩为

$$\begin{aligned}\text{均值} &= 113.51 - 100.24 = 13.27,\\ \text{二阶矩} &= 45\ 494.83 - 23\ 993.47 - 2(500)(13.27) = 8\ 231.36,\\ \text{方差} &= 8\ 231.36 - 13.27^2 = 8\ 055.27,\\ \text{协方差} &= \frac{\sqrt{8\ 055.27}}{13.27} = 6.76.\end{aligned}$$

尾部加重使得风险增加. 全年的 3 个值分别为 278.67, 172 858.56, 1.49.

14.4.3 第二个模型

由图 14-5 可知如果选择单参数模型, 则应该考虑非零众数. 因为数据都在 1 000 左右, 区间可以设为 0.5~5.5, 5.5~10.5 等等. 可以考虑对数正态、Weibull、gamma 和混合模型 (包括指数分布) 等分布, 对数正态分布明显是最合适的 (利用 SBC). 参数为 $\hat{\mu} = 4.062\ 6, \hat{\sigma} = 1.146\ 6$. 卡方拟合优度检验 (将超过 200 的观测值单放在一组) 的统计量为 7.77, p 值为 0.73. 图 14-6 将对数正态模型和经验模型进行了比较, 说明拟合效果很好.

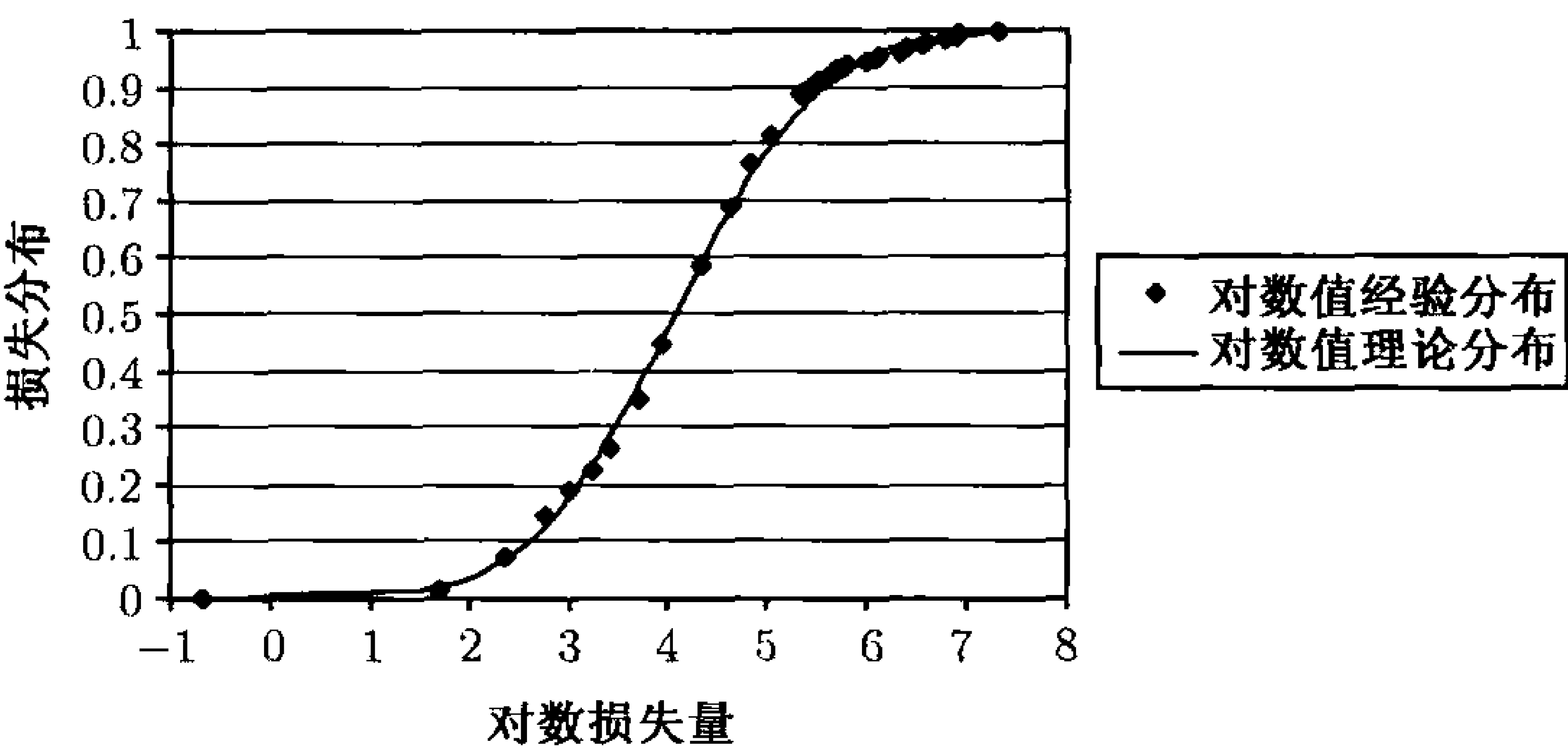


图 14-6 分布函数图

14.5 总损失实例 I

本节的例子是对目前介绍的很多方法的综合. 这里讨论的承保责任可能比实际中的要复杂, 但可以帮助我们考虑解决各方面的问题.

例 14.1 一个咨询精算师正在为某个团体医疗保险进行定价. 目标是确定保险公司支出的期望. 保单条款 (对于每个雇员) 如下:

(1) 对于雇员或其家属的每次就医, 雇员本人支付前 500 元和超过 50 500 的部分. 即对于每次就医, 保险人最多赔付 50 000;

(2) 在每个公历年, 雇员自己的总支出不超过 1 000 的免赔额, 但对损失超过 50 500 的部分没有限制;

(3) 每次就医的时间是指去医院就诊时的公历年, 即使就医时间延迟到下一年, 所有赔付也对应于最初就诊的那一年.

(4) 无论家属人数如何, 保费都是相同的.

经验数据如表 14-6 和表 14-8. 表 14-7 是雇员家属人数的数据.

表 14-6 住院就医次数/人, 年

每个家庭成员的就医次数	家庭成员人数
0	2 659
1	244
2	19
3	2
4 或更多	0
总计	2 924

第一步利用 3 个数据集来确定参数模型. 有 12 个分布适合表 14-6 的数据, 最佳的单参数分布是负对数似然 (NLL) 为 969.251 的几何分布, 并且卡方拟合优

度检验的 p 值为 0.532 5. 最佳的两参数模型是零点修正的几何模型, NLL 改进为 969.058, 但由似然比检验并不足以判断第二个参数的适用性. 最佳的三参数模型是零点修正的负二项几何模型, NLL 为 969.056, 也不足以将几何分布排除在我们的选择之外. 因为两参数模型和三参数模型没有足够的自由度进行卡方检验, 我们选择 $\beta = 0.098\ 49$ 的几何分布.

表 14-7 每个雇员的家庭成员数

雇员的家庭成员数	雇员数
1	84
2	140
3	139
4	131
5	73
6	42
7	27
8 或更多	33
总计	669

表 14-8 每次就医的损失量

每次就医的损失量	就医次数
0~250	36
250~500	29
500~1 000	43
1 000~1 500	35
1 500~2 500	39
2 500~5 000	47
5 000~10 000	33
10 000~50 000	24
50 000~	2
总计	288

对于表 14-7 的数据, 只有零点截断的分布可行. 最佳的单参数模型是零点截断 Poisson 分布, NLL 为 1 298.725, p 值接近零. 两参数零点截尾负二项分布的 NLL 为 1 292.532, 有明显的改进, p 值为 0.257 1, 说明这是一个可接受的选择, 参数为 $r = 13.207, \beta = 0.258\ 84$.

对于表 14-8 的数据, 有 15 种连续分布可选. 给定参数个数的最佳 4 个分布列在表 14-9 中, 明显看出 Pareto 分布是最佳选择, 参数为 $\alpha = 1.669\ 3, \theta = 3\ 053$.

其余的计算可以用递推式得到, 逆变换或者模拟的办法也可以达到同样的效果.

表 14-9 每次就医的损失量的四个最优拟合模型

名称	参数个数	NLL	p 值
逆指数	1	632.632	接近 0
Pareto	2	601.642	0.981 8
Burr	3	601.612	0.947 6
转移 beta	4	601.553	0.879 8

在确定每个家庭成员的赔付分布时, 第一步要考虑免赔. 若频率分布为几何分布, 而个体损失分布为 Pareto 分布, 则最高免赔额为 500, 即任何超过 500 的损失都取定为 500. 在离散化递推计算时, 应该对 500 以内等分区间, 然后对 500 内无法解释的概率都放在 500 这个点. 本例中的区间间隔为 1, 表 14-10 给出了最前面和后面的几个离散分布的值. 利用递推公式, 容易得到 3 000 以上的概率非零. 然而, 考虑到雇员的总免赔额, 1 000 以上的赔付是没有影响的. 表 14-11 给出了一些

表 14-10 限制为 500 的离散化 Poisson 分布

损失	概率
0	0.000 273
1	0.000 546
2	0.000 546
3	0.000 545
⋮	⋮
498	0.000 365
499	0.000 365
500	0.776 512

表 14-11 每家庭成员总免赔额的概率

损失	概率
0	0.910 359
1	0.000 045
2	0.000 045
3	0.000 045
⋮	⋮
499	0.000 031
500	0.063 386
501	0.000 007
⋮	⋮
999	0.000 004
1 000	0.004 413
1 001	0.000 001
⋮	⋮

这样的概率.

接下来得到雇员每年总免赔额的分布, 这也是一个复合分布. 频率分布为截断的负二项分布, 个体损失分布是已经得到的家庭成员损失分布. 也可以用递推式得到这个分布. 因为免赔限额为 1 000, 所有在 1 000 以上的概率都放在 1 000. 表 14-12 给出了该总体分布的一些概率值. 注意到免赔超过 1 000 的赔付概率很小, 保险人因为限制了被保险人自己承担的成本而花费的成本也很小. 利用离散分布, 容易得到总免赔的均值和标准差, 分别为 150.02 和 274.42.

表 14-12 每个雇员总免赔额的概率分布

损 失	概 率
0	0.725 517
1	0.000 116
2	0.000 115
3	0.000 115
⋮	⋮
499	0.000 082
500	0.164 284
501	0.000 047
⋮	⋮
999	0.000 031
1 000	0.042 343

接下来要计算个体损失低于上限 50 000 的保险总成本的期望值, 这可以通过分析直接得出. 对于 Pareto 分布, 每次损失的期望支出为 $\mathrm{E}(X \wedge 50\ 500) = 3\ 890.87$. 每个家庭成员的期望损失次数为参数 0.098 49 的几何分布的均值. 每个雇员的家庭成员个数服从零点截断负二项几何分布, 期望值为 3.590 15. 因此每个雇员的期望损失次数为 0.098 49(3.590 15)=0.353 612. 则每个个体的期望损失值为 0.353 612 (3 890.87)=1 375.86.

保险公司期望成本为 1 375.86−150.02=1 225.84. 最后要注意的是, 除了随机模拟的方法, 没有其他的方法可以得到保险人赔付的概率分布. 同例 17.7 的情况类似, 很容易得到整体的分布和被保险人的分布 (这种情况下, 如果损失超过 50 500 的支付可以忽略), 但不能得到保险公司的损失分布. □

14.6 总损失实例 II

已知通过认真地建模得到个体损失服从对数正态分布, 参数为 $\mu = 10.543$, $\sigma = 2.313\ 15$. 还可以确定损失次数的分布为 Poisson 分布, 参数为 $\lambda = 0.015\ 457\ 8$.

考虑一个超额损失再保险, 再保方支付超过免配额 d, 最大为 $u-d$ 的支付, 其中 u 为没有免赔额 d 时的赔付上限. 有两种构造再保支付分布的方法. 第一种方法是计算每次支付额的分布, 基于这种方法, 严格的分布函数为混合型, 其概率分布函数为

$$f_Y(x)=\frac{f_X(x+d)}{1-F_X(d)},\quad 0\leqslant x<u-d,$$

离散概率为

$$\Pr(Y=u-d)=\frac{1-F_X(u)}{1-F_X(d)}.$$

这个分布可以利用递推公式或 FFT 进行离散计算, 或是利用 Heckman-Meyers 方法用直方图近似. 另外, 频率分布必须能够反映分布是赔付次数的分布而不是损失次数的分布, 所以新的 Poisson 参数为 $\lambda[1-F_X(d)]$.

14.6.1 单个保单的分布

考虑对于不同的 d 与 u 组合后每张保单的损失分布. 用 Poisson 参数代表不同的组合, 而且在 10 000 之内进行离散化取整处理, 递推计算. 在所有情况下, 90%和99%分位数都是零, 说明多数情况下超额损失再保险的支付为零. 这并不奇怪, 因为不发生损失的概率为 exp(−0.015 457 8)=0.985, 而有免赔时的概率会更高. 对于 d 与 u 的不同组合, 均值、标准差和变异系数都列在表 14-13 中.

表 14-13 单个保单的超额再保

免赔 (10^6)	限额 (10^6)	均 值	标准差	C.V.
0.5	1	778	18 858	24.24
0.5	5	2 910	94 574	32.50
0.5	10	3 809	144 731	38.00
0.5	25	4 825	229 284	47.52
0.5	50	5 415	306 359	56.58
1.0	5	2 132	80 354	37.69
1.0	10	3 031	132 516	43.72
1.0	25	4 046	219 475	54.24
1.0	50	4 636	298 101	64.30
5.0	10	899	62 556	69.58
5.0	25	1 914	162 478	84.89
5.0	50	2 504	249 752	99.74
10.0	25	1 015	111 054	109.41
10.0	50	1 605	205 939	128.71

很自然地风险程度 (由变异系数度量) 会随免赔额或限额的增加而增加, 而且只出售一张保单的风险是极大的.

14.6.2 100 个保单–超额损失保单组

接下来考虑 100 个再保保单的组合. 如果假设其具有相同的免赔额和限额, 则总分布只需改变频率分布. 100 个独立的 Poisson 变量之和仍服从 Poisson 分布, 且参数为原来参数的 100 倍. 重复上述过程并修改 Poisson 参数, 结果如表 14-14.

表 14-14 100 个保单的超额再保

免赔 (10^6)	限额 (10^6)	均值 (10^3)	标准差 (10^3)	C.V.	分位点 (10^3)	
					90	99
0.5	5	291	946	3.250	708	4 503
0.5	10	381	1 447	3.800	708	9 498
0.5	25	482	2 293	4.752	708	11 674
1.0	5	213	804	3.769	190	4 002
1.0	10	303	1 325	4.372	190	8 997
1.0	25	405	2 195	5.424	190	11 085
5.0	10	90	626	6.958	0	4 997
5.0	25	191	1 625	8.489	0	6 886
10.0	25	102	1 111	10.941	0	1 854

因为保单相互独立, 均值是原来的 100 倍, 标准差是单个保单的 10 倍, 这表明变异系数为原来的 1/10. 在所有情况中, 99% 分位数均大于零, 这表明风险增加了, 而在实际中它表明索赔支付可能性更大.

14.6.3 100 个保单–总损失止损处理

现在考虑对总损失的再保险. 假设保单没有免赔额但有上限 u. 同样 100 个保单, 但再保支付为总损失超过免赔额 a 的部分. 对于给定的限额, 如前面对损失量分布的修正, Poisson 参数要乘上 100, 然后用一些算法得到总损失分布. 令其分布的累积分布函数为 $F_S(s)$, 或在离散化分布的情况下 (由递推法则或是 FFT 得到的结果), 概率函数为 $f_S(s_i), i=1,\cdots,n$. 免赔额为 a 的再保险分布函数 S_r 为

$$\begin{aligned} F_{S_r}(s) &= F_S(s+a), \quad s \geqslant 0, \\ f_{S_r}(0) &= F_S(a) = \sum_{s_i \leqslant a} f_S(s_i), \\ f_{S_r}(r_i) &= f_S(r_i+a), \quad r_i = s_i - a, \quad i=1,\cdots,n. \end{aligned}$$

用通常的方法可以得到矩和百分位数.

在 10 000 之内利用递推公式, 得到不同止损免赔和个体限额情况下的结果, 见表 14-15. 结果与超额损失赔付相似. 对于多数情况, 随着个体损失限额或总体免赔的增加, 由变异系数度量的风险都会增加. 一个例外是当限额和免赔都是 5 000 000 的情况, 这种设置风险很大, 因为这是唯一在再保生效之前发生两个损失

的情况.

表 14-15 100 个保单的总止损再保

免赔 (10^6)	限额 (10^6)	均值 (10^3)	标准差 (10^3)	C.V.	分位点 (10^3)	
					90	99
0.5	5	322	1 003	3.11	863	4 711
0.5	10	412	1 496	3.63	863	9 504
0.5	25	513	2 331	4.54	863	11 895
1.0	5	241	879	3.64	363	4 211
1.0	10	331	1 389	4.19	363	9 004
1.0	25	433	2 245	5.19	363	11 395
2.5	5	114	556	4.86	0	2 711
2.5	10	204	1 104	5.40	0	7 504
2.5	25	306	2 013	6.58	0	9 895
5.0	5	13	181	13.73	0	211
5.0	10	103	714	6.93	0	5 004
5.0	25	205	1 690	8.26	0	7 395

现假设这 100 个保单的 Poisson 参数不同 (但有相同的索赔额分布). 假设其中 30 个的参数为 $\lambda = 0.016\ 224\ 9$, 因此这 30 个索赔构成的组合服从 Poisson 分布, 均值为

$$30(0.016\ 224\ 9) = 0.486\ 747.$$

第二组 (50 个) 的 Poisson 参数为 50(0.017 408 7)=0.870 435. 第三组 (20 个) 的 Poisson 参数为 20(0.009 612 1)=0.192 242. 有 3 种方法可以构造这 3 组之和的分布.

(1) 因为独立 Poisson 分布的和也是 Poisson 分布, 所以损失总次数服从 Poisson 分布, 其 Poisson 参数为 1.549 424. 一般损失量分布还是对数正态分布. 简化成一个复合分布, 可以用任意方法估计.

(2) 分别得到 3 个聚合分布. 如果利用递推或是 FFT 算法得到 3 个离散分布, 可以用卷积得到和的分布.

(3) 如果用 FFT 或是 Heckman-Meyers 算法, 可以得到 3 个变换然后再相乘. 给出乘积后的逆变换形式.

每种方法都各有优缺点. 第一种方法要求其和具有已知形式的频率分布. 如果索赔额分布不同, 就无法将其合并形成一个模型. 它的主要优点是, 当模型适用时只需进行一次聚合计算.

第二种方法的优点在于它对于频率分布和损失程度分布没有限制. 缺点是计算机存储量的快速膨胀, 例如, 第一个分布需要 3 000 个点, 第二个分布 5 000 个点,

第三个分布 2 000 个点 (3 个分布的离散化区间相同), 则联合分布将有 10 000 个点. 本节的最后还会进一步对此说明.

第三种方法对于每个分离的模型也没有要求. 它的缺点和第二个相同, 但是存储的膨胀是在合并之前发生的, 即 3 个分布都要计算 10 000 个点. 似乎没有办法避免这一点.

14.6.4 数值卷积计算

剩下的问题是进行数值卷积计算时需要的点数. 当个体分布产生大量的离散点时, 计算机的存储能力会成为障碍, 从而产生问题. 下例是这个问题的小规模版本, 介绍了一种简单解法.

例 14.2 下面给出了 2 个离散分布的概率函数. 计算机存储的最大向量长度为 6. 近似确定两个随机变量和的概率函数.

x	$f_1(x)$	$f_2(x)$
0	0.3	0.4
2	0.2	0.3
4	0.2	0.2
6	0.2	0.1
8	0.1	0.0

解 2 个随机变量和的最大可能值为 14, 需要长度为 8 的向量进行存储. 利用卷积方法得到下面结果.

x	0	2	4	6	8	10	12	14
$f(x)$	0.12	0.17	0.20	0.21	0.16	0.09	0.04	0.01

若只有 6 个点可用, 则区间跨度必须增加为 14/5=2.8. 为此需要逆向插值, 将每一个非 2.8 倍数的点上的概率分配到离它最近的两个 2.8 倍数的点上. 例如, 在 x=8 处的概率为 0.16, 要将它分配在 5.6 和 8.4 上. 因为 8 处于从 5.6 到 8.4 的线段中 2.4/2.8 的位置, 因此有 6/7 的概率分配到 8.4 上, 剩余 1/7 的概率分配到 5.6 上. 整个分配过程如表 14-16. 当概率被分配到每个 2.8 倍数的点后, 就可以构造和分布的近似结果. 近似分布如下.

x	0	2.8	5.6	8.4	11.2	14.0
$f(x)$	0.168 6	0.235 7	0.288 6	0.205 7	0.080 0	0.021 4

这种方法保持全概率为一和均值不变 (真实分布与近似分布的均值都为 5.2).

□

有一种改进方法可以消除部分的存储需求, 当一个分布需要较大的向量进行存储时尾概率通常很小. 对它进行卷积计算时, 下一步向量的尾概率可能很小以至可

以忽略. 因此可以不保留这些点, 从而不增加存储的问题.

表 14-16　例 14.2 的概率分配过程

x	$f(x)$	下限	概率	上限	概率
0	0.12	0	0.120 0		
2	0.17	0	0.048 6	2.8	0.121 4
4	0.20	2.8	0.114 3	5.6	0.085 7
6	0.21	5.6	0.180 0	8.4	0.030 0
8	0.16	5.6	0.022 9	8.4	0.137 1
10	0.09	8.4	0.038 6	11.2	0.051 4
12	0.04	11.2	0.028 6	14.0	0.011 4
14	0.01	14.0	0.010 0		

还有很多其他的技巧. 附录中的 Bailey[9] 介绍了一种方法可以保留前三阶矩, 它还提供了消除或组合小概率存储分配的指导方法.

综合习题

本节的习题同本章前面的例题类似. 这些问题都是基于已发表的论文而提出的.

14.2　在纽约州有一项特殊的基金, 专为工伤险的小概率事件进行赔偿, 其中一种情况是那些重新开庭的案件. Hipp[77] 收集了从事故发生到案件重新审理的时间数据, 包括在 1933 年 4 月 24 日至 1936 年 12 月 31 日之间的重新审理案件, 数据如表 14-17 所示. 试为从案发到重新审理的时间间隔建立一个参数模型. 根据定义, 索赔必须经过 7 年的时间才可以申请重新开庭, 因此这个模型应该考虑大于等于 7 年的这个条件.

表 14-17　例 14.2 工伤险案件重新审理的时间

年	重新审理数	年	重新审理数
7~8	27	15~16	13
8~9	43	16~17	9
9~10	42	17~18	7
10~11	37	18~19	4
11~12	25	19~20	4
12~13	19	20~21	1
13~14	23	21+	0
14~15	10		
		总计	264

14.3　在 Arthur Bailey[6] 最初于 1942 年和 1943 年发表的两篇论文中, 曾在第 1 篇论文 [6] 第 51 页写道"采用抽样分布的另一个好处是为免赔和超赔进行费率设计". 在第 2 篇论

文 [7] 中, 他列出了一些损失率分布的数据 (表 14-18). 在该论文中, 他认为对数正态模型可以很好地拟合数据, 并通过了卡方检验. 试判断这点是否正确? 是否有更好的模型?

表 14-18 损失率数据–习题 14.3

损失比	个数
0.0~0.2	16
0.2~0.4	27
0.4~0.6	22
0.6~0.8	29
0.8~1.0	19
1.0~1.5	32
1.5~2.0	10
2.0~3.0	13
3.0+	5
总计	173

14.4 在 1979 年 Hewitt and Lefkowitz[56] 观测了机动车车身责任险的损失数据 (表 14-19), 得出结论采用 gamma 分布和对数 gamma 分布的二元混合分布 [X 服从 gamma 分布, 则 Y=exp(X) 的分布为对数 gamma 分布] 优于对数正态分布. 你是否同意这个结论? 并考虑 gamma 分布与对数 gamma 分布的对比.

表 14-19 机动车车身责任损失–习题 14.4

损失量	个数	损失量	个数
0~50	27	750~1 000	8
50~100	4	1 000~1 500	16
100~150	1	1 500~2 000	8
150~200	2	2 000~2 500	11
200~250	3	2 500~3 000	6
250~300	4	3 000~4 000	12
300~400	5	4 000~5 000	9
400~500	6	5 000~7 500	14
500~750	13	7 500~	40
		总计	189

14.5 Patrik 在 1980 年的论文 [102] 中讨论了本书中的很多方法. 其中一个例子是美国保险服务局 (ISO) 公布的雇主、农场主和佃户的人身伤害责任险的数据. 现有两种不同赔付限额的保单, 都是 1976 年签单并且损失发生截至 1978 年底的保单. 表 14-20 中的分组已经对论文中的数据进行了压缩. 是否可以用同一个模型 (参数相同或不同) 描述两种赔付限额保单?

表 14-20 OLT 人身意外伤害责任险损失–习题 14.5

损失量 (10^3)	300 限额	500 限额	损失量 (10^3)	300 限额	500 限额
0~0.2	10 075	3 977	11~12	56	22
0.2~0.5	3 049	1 095	12~13	47	23
0.5~1	3 263	1 152	13~14	20	6
1~2	2 690	991	14~15	151	51
2~3	1 498	594	15~20	151	54
3~4	964	339	20~25	109	44
4~5	794	307	25~50	154	53
5~6	261	103	50~75	24	14
6~7	191	79	75~100	19	5
7~8	406	141	100~200	22	6
8~9	114	52	200~300	6	9
9~10	279	89	300~500	10[a]	3
10~11	58	23	500~		0
			总计	24 411	9 232

a 考虑 300 以上的损失量.

14.6 表 14-21 中的数据是 Fisher[37] 收集的美国 1910 年之前 25 年的矿灾数据. 这里的矿灾定义为造成 5~9 名矿工索赔的矿难. Fisher 认为 Poisson 分布是一个很好的模型. 这个结论是否正确? 是否有更好的模型?

表 14-21 矿难数据–习题 14.6

灾难数	年数	灾难数	年数
0	1	7	3
1	1	8	1
2	3	9	0
3	4	10	1
4	5	11	1
5	2	12	1
6	2	13+	0

14.7 Harwayne[49] 主要讨论驾驶违规记录和事故数之间的关系. 他记录了加利福尼亚州驾驶员违规数的数据, 表 14-22 中每一栏的 6 个数据, 用负二项分布拟合是否合理? 如果是这样, 参数相同是否合理? 是否能得到期望事故数随违规数增加而增加的结论?

表 14-22 事故数和违规数–习题 14.7

事故数	违规数					
	0	1	2	3	4	5+
0	51 365	17 081	6 729	3 098	1 548	1 893
1	3 997	3 131	1 711	963	570	934
2	357	353	266	221	138	287
3	34	41	44	31	34	66
4	4	6	6	6	4	14
5+	0	1	1	1	3	1

14.8 在 1961 年 Simon[122] 提出了零点修正的负二项分布, 数据为俄勒冈州各种高速公路路段上在 1 英里内的年事故数, 数据见表 14-23. Simon 称零点修正负二项分布优于负二项分布. 他的结论是否正确? 是否有更好的模型?

表 14-23 交通事故数–习题 14.8

意外事故数	路段数	意外事故数	路段数
0	99	6	4
1	65	7	0
2	57	8	3
3	35	9	4
4	20	10	0
5	10	11	1

第五部分　统计估计的调整及随机模拟

第 15 章 插值与平滑

15.1 引 言

本教材之前讨论的建模方法主要是基于概率统计学, 认为数据来自有概率分布的样本空间. 所讨论的估计量也都是概率分布的函数, 例如: 概率密度函数, 累积分布函数, 风险率 (死亡率), 均值, 方差等.

与之对比, 本章所要描述的方法则来源于数值分析, 而没有在概率统计的概念上加以特别的考虑.

在实际中, 很多这样的数值方法已经陆续在概率和统计的框架之下采用. 虽然它们的基本理念易于理解, 但是大多数方法需要大量的计算, 因此往往需要计算机编程技术. 本章中讨论的技术仅仅是这种复杂问题中最基本的部分.

我们的目标是通过某种特定的评判准则将一系列的数据拟合成一条平滑的曲线. 这种方法在精算学以及其他领域中都有着很多应用. 假设在平面上有一个点集, 在现实中它代表某个变量的一系列连续的观测, 例如, 它可以是月通货膨胀率的一系列连续观测, 或者是平均年索赔成本的一系列连续观测, 也可能是各年龄死亡率的一系列连续观测. 本章采用非参数的方法, 也就是说考虑的模型并不仅仅限定在用几个参数表示的简单数学函数中. 本章采用的方法不对最终曲线的形状给出很多限制. 在当最终曲线的形状很复杂时这种方法尤其有用.

例如考虑将某人群在短期内的死亡概率 (例如 q_x 函数) 的曲线. 由于受到初生婴儿死亡的影响, 死亡概率在幼年时期下降得非常快, 然后在 10 岁左右的区间内相对平缓, 在十几岁的区间内死亡率有缓慢的增长, 在 18~25 岁之间由于意外事故的影响会出现先增后减的情况 (尤其是对男性而言), 其后死亡率增加缓慢但在更高年龄加速增长. 像这样的曲线是无法用一个简单函数 (尽管有的简单函数也有 8 个或者更多的参数) 进行表示的.

历史上, 将平滑一个不规律的观测点集的过程称作修匀. 典型的这类点集是一年期死亡率或者是一些其他的偶然事件 (例如残疾, 失业或者事故) 的发生概率. 而本章中描述的模型将不仅仅限于上述几种应用. 事实上, 它们可以应用于任何连贯的观测点集.

在修匀理论中, 假设存在某种无法直接观测的真实的曲线或者函数, 因此我们只能通过估计或者近似给出其真实的形式. 修匀理论需要在较高的拟合度与平滑

度之间进行平衡, 拟合度往往体现于用 "噪声" 曲线例如高阶多项式来拟合数据, 而平滑度则意味着一条简单的曲线例如直线或者指数分布曲线.

早期的精算教科书 (例如 Miller[94]) 有很多这类古典的方法. 其中包括简单的图表方法. 此法是用一块工程学绘图中采用的特殊的法式曲尺, 或者是使用样条和砝码来绘图的. 法式曲尺是一块带有光滑外边缘的木头, 外边缘有着渐变的直径. 这样就可以通过特定的点来画出特定的曲线. 而样条是指一个有弹性的金属或塑料制的薄杆. 早期工程师制图时, 把这样的薄杆 (样条) 固定在样点上, 在其他地方让它自由弯曲, 然后画下长条的曲线, 称这样的曲线为样条曲线. 这是一种很自然地用来推导结构形状的方法, 能使其拥有最大的强度. 精算师们发展了这种方法. 他们考虑了基于移动平均的数学方法, 基于插值的方法以及基于寻找一个拟合与平滑之间的平衡的方法. 这些方法都产生于 20 世纪的早期, 有些甚至更早. 主要是有限差分方法, 经常将 4 阶以及更高阶的差分直接设为零, 这也间接的导致了三阶多项式的使用. 人们同时推导出了包含差分的方程, 这样精算师们只要通过纸与铅笔就可以推导出平滑函数. 这些方程的出现远远早于计算器的出现, 而在这之后很长时间计算机才被人类制造出来. 关于此类方法以及一些新改良变化的更多新总结可以在 London[84] 中找到.

随着计算机技术在 20 世纪 50 年代和 20 世纪 60 年代的发展, 很多依赖计算机的数学程序被开发出来. 其中包括了样条理论, 而这时的样条不再是现实中的器械. 在渐进理论中, 样条的目的是找到一个拟合与平滑之间的合适的平衡. 可以很容易地在电脑上通过线性方程组来解出它的值. 现代的样条理论可以追溯到 Schoenberg[117].

本章仅仅着眼于现代样条理论中的插值和平滑技术. 这些现代技术非常强大灵活, 已经远远地超过了旧的方法.

15.2 多项式插值与平滑

现有 $n+1$ 个点: $(x_0, y_0), (x_1, y_1), \cdots, (x_n, y_n)$, 其中 $x_0 < x_1 < x_2 < \cdots < x_n$, 则存在唯一的 n 阶多项式恰通过这 $n+1$ 个点. 称这个多项式为**配置**多项式, 可以表示为

$$f(x) = \sum_{j=0}^{n} a_j x^j, \tag{15.1}$$

满足

$$f(x_j) = y_j, \quad j = 0, 1, \cdots, n. \tag{15.2}$$

方程组 (15.2) 由 $n+1$ 个方程与 $n+1$ 个未知数 $\{a_j; j = 0, 1, \cdots, n\}$ 构成. 当 n 很大时, 要求出这个方程组的数值解将会比较困难.

不过, 我们可以不用求解上述方程组而是由 $n+1$ 个点的纵坐标给出解的解析表达式. 由拉格朗日公式

$$
\begin{aligned}
f(x) =& y_0 \frac{(x-x_1)(x-x_2)\cdots(x-x_n)}{(x_0-x_1)(x_0-x_2)\cdots(x_0-x_n)} \\
&+ y_1 \frac{(x-x_0)(x-x_2)\cdots(x-x_n)}{(x_1-x_0)(x_1-x_2)\cdots(x_1-x_n)} \\
&+ \cdots \\
&+ y_n \frac{(x-x_0)(x-x_1)\cdots(x-x_{n-1})}{(x_n-x_0)(x_n-x_1)\cdots(x_n-x_{n-1})} \\
=& \sum_{j=0}^{n} y_j \frac{(x-x_0)\cdots(x-x_{j-1})(x-x_{j+1})\cdots(x-x_n)}{(x_j-x_0)\cdots(x_j-x_{j-1})(x_j-x_{j+1})\cdots(x_j-x_n)}. \qquad (15.3)
\end{aligned}
$$

(15.3) 式的每一项都是一个 n 阶多项式, 而对每个 $j=0,1,\cdots,n$, 当 $x=x_j$ 时 (15.3) 右式的取值即为 y_j. 这正说明 (15.3) 式就是一个配置多项式.

这个 n 阶多项式 $f(x)$ 在 (x_0,y_0) 到 (x_n,y_n) 之间进行的插值, 通过了所有的点 $\{(x_j,y_j);j=1,\cdots,n-1\}$. 然而, 当 n 比较大时 $f(x)$ 的形状将呈现出很大的上下摆动. 尤其是当原序列 $\{(x_j,y_j);j=0,\cdots,n\}$ 存在某些 "噪声" 的时候, 这些噪声可能是由于测量误差或者随机波动导致的.

例 15.1 表 15-1 的数据摘自 Miller[94] 第 62 页. 它们是以 5 年为一组的死亡率. 这时的死亡率是由索赔金额与总风险保费之比估计得到的[①]. 图 15-1 给出了这些死亡率估计的变化.

表 15-1 例 15.1 的死亡率

j	年 龄	风险暴露	实际死亡赔付	每千人估计死亡率
0	25~29	35 700	139	3.89
1	30~34	244 066	599	2.45
2	35~39	741 041	1 842	2.49
3	40~44	1 250 601	4 771	3.81
4	45~49	1 746 393	11 073	6.34
5	50~54	2 067 008	21 693	10.49
6	55~59	1 983 710	31 612	15.94
7	60~64	1 484 347	39 948	26.91
8	65~69	988 980	40 295	40.74
9	70~74	559 049	33 292	59.55
10	75~79	241 497	20 773	86.02
11	80~84	78 229	11 376	145.42
12	85~89	15 411	2 653	172.15
13	90~94	2 552	589	230.80
14	95~	162	44	271.60
	共计	11 438 746	220 699	

① 死亡与风险暴露都是以 1 000 美元作为单位. 在死亡率的研究中, 由于对大额保单将赋予更大的权重, 所以经常以金额而不是人数作为计算依据. 表中的死亡率是死亡赔付与风险暴露的比值. 由于四舍五入, 最后一个数据与原数据略有偏差.

解 这里的死亡率是最大似然估计, 并假设各年龄服从相互独立的二项分布. 注意到估计过程中的很大变化. 当然, 我们希望死亡率随着年龄的变化是相对平滑的. 图 15-1 将估计的死亡率点用直线连接, 而图 15-2 是用配置多项式来拟合观测点. 可以发现, 图 15-2 表现出剧烈起伏, 尤其在尾部出现了巨大的摆动. □

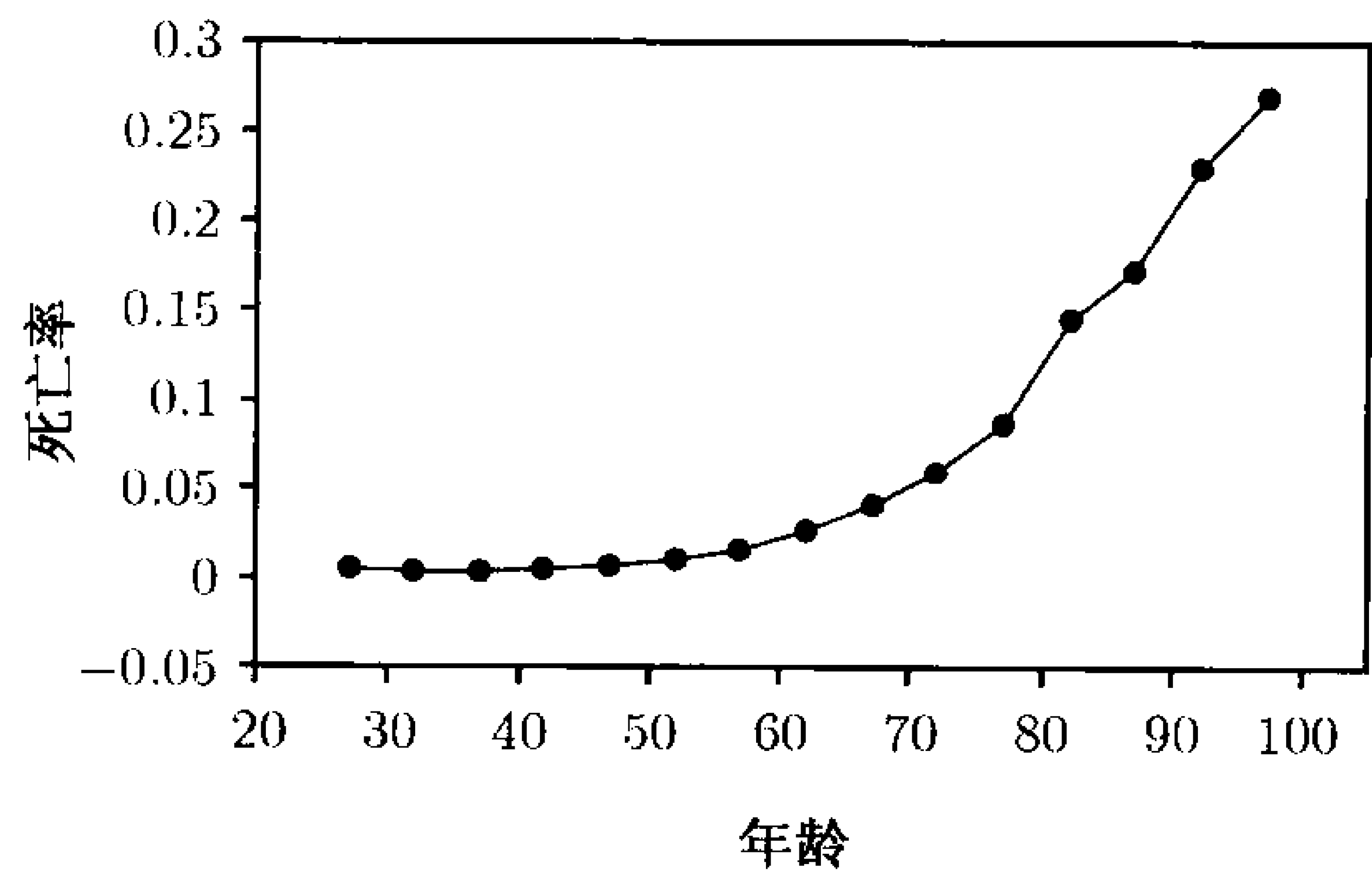


图 15-1 例 15.1 的死亡率

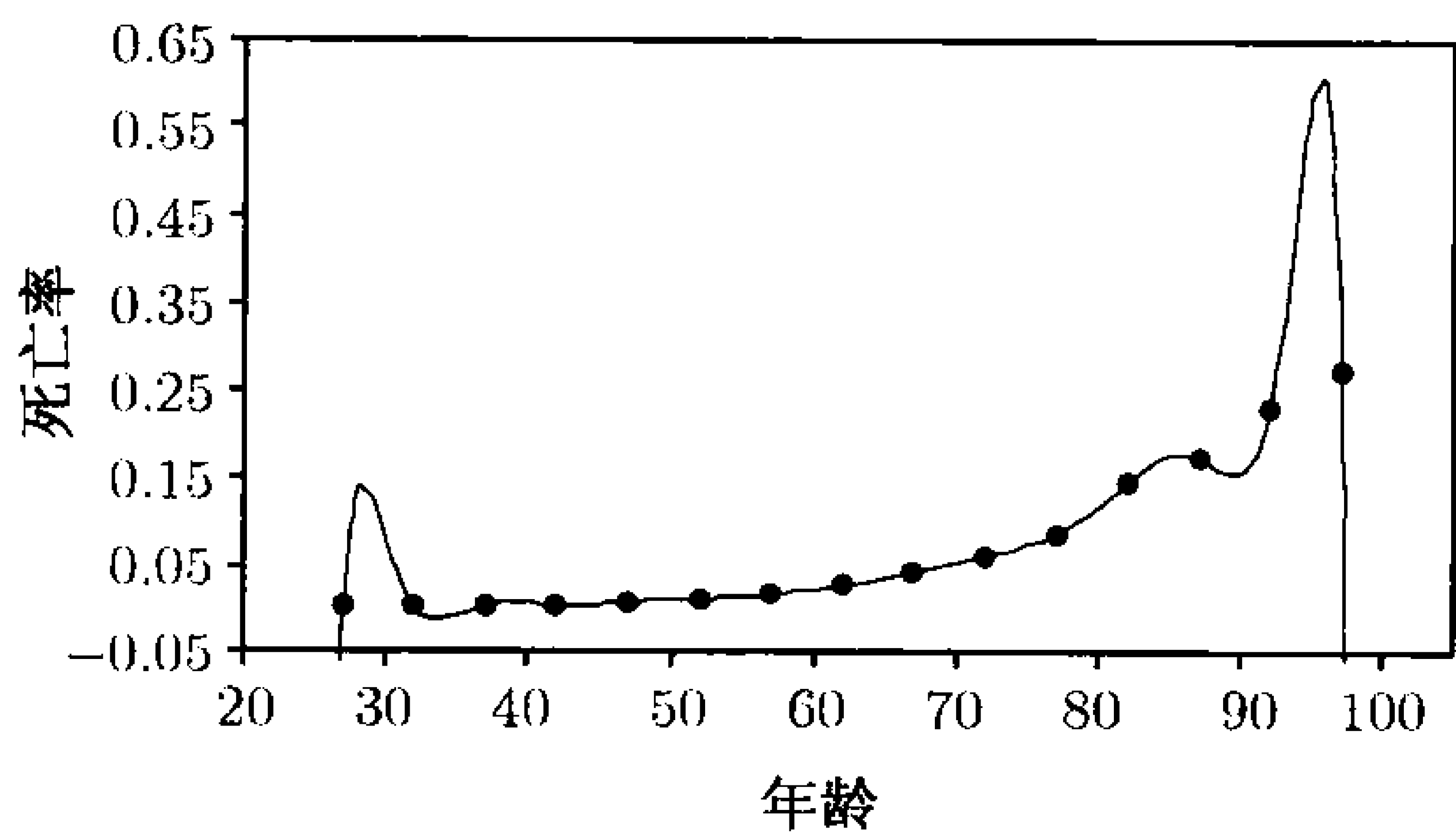


图 15-2 死亡率数据的配置多项式

为了避免出现这种过度摆动或起伏现象, 有时也会采用较低阶的多项式进行插值. 例如, 有时候就是用直线进行点连接. 然而, 由于在连接点处一阶导不存在, 这种方法会生成一个锯齿形的序列.

另一种方法是将一系列低阶多项式缝合起来. 例如, 可以分别对连接点 $(x_0, x_1, x_2), (x_2, x_3, x_4)\cdots$ 建立二次多项式插值. 不过这时类似上面用直线将相邻点链接起来的情况, 此时在节点 $x_2, x_4, \cdots$ 处不平滑, 左右两侧的斜率与曲率不一致. 一种修正的方法是改变这些点在左侧与右侧的导数, 强制它们相同. 这样显然保证了在节点处为一个平滑的过程, 而这也是样条方法中的关键理念. 插值样条是由一组分段的多项式函数缝合而成, 这些函数首先应穿过节点, 同时由其他约束使得它们在结合点处保持平滑. 为了减少起伏, 令多项式的阶数尽量较低. 15.3 节中将讨论三次样条插值.

还有一种插值的方法是平滑，更准确地说是对观测数据拟合一个平滑的函数，但是并不要求该函数穿过所有的观测点. 多项式本身在形状上具有很大的灵活性，然而，这种形状上的灵活性经常使得用多项式进行外推时有一定的风险，尤其是对于高阶多项式而言. 图 15-2 就是这样的情况：如果通过图中的曲线外推，尽管仅仅是向后外推一年，得到的结果也是完全不可信的. 正如在本书之前所讨论的其他模型的拟合问题一样，我们需要为模型拟合建立一个评判标准. 对于多项式平滑采用最小均方标准. 图 15-3 到图 15-6 展示了对例 15.1 的数据进行的 2, 3, 4, 5 阶的多

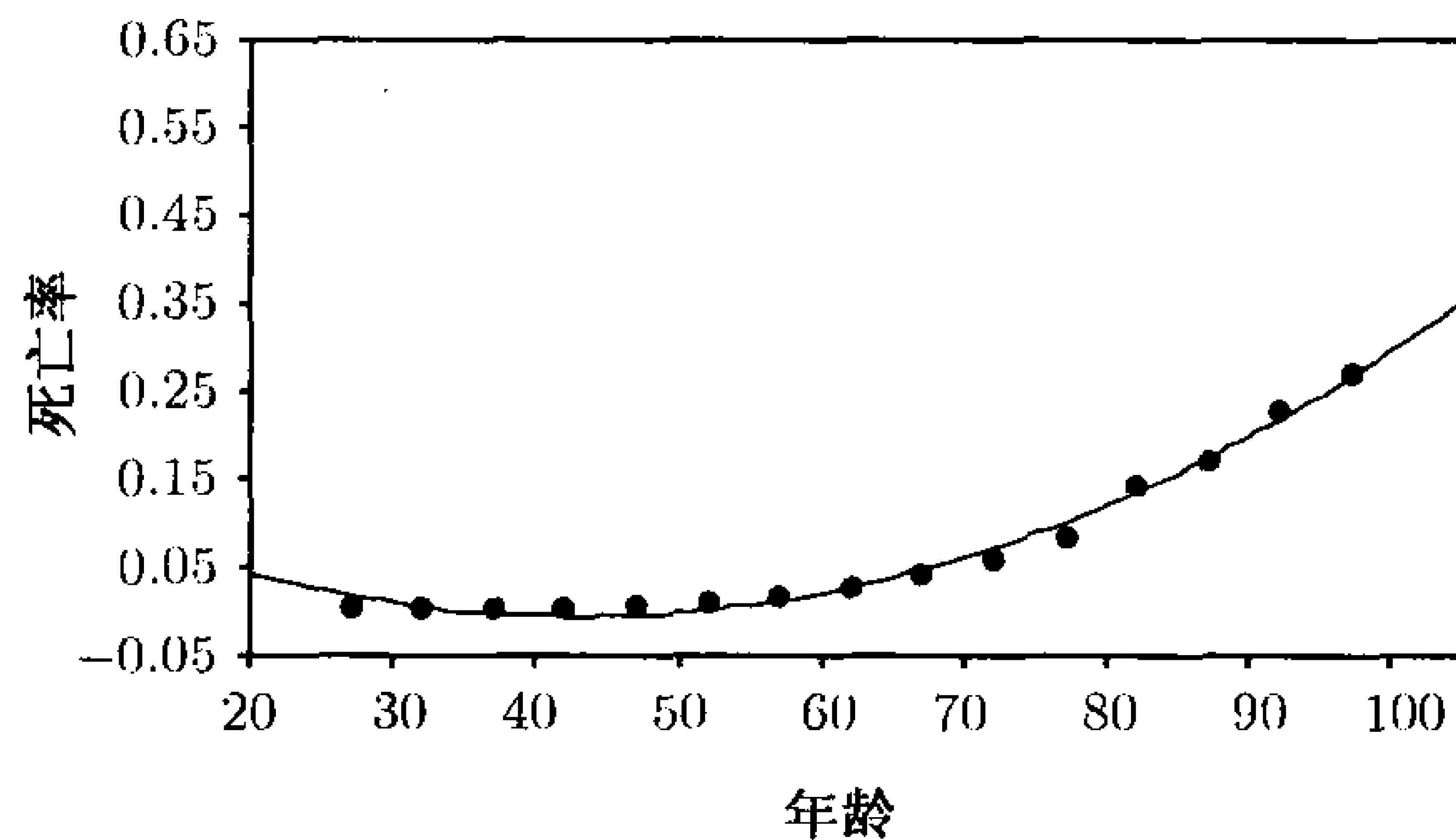


图 15-3 二阶多项式拟合

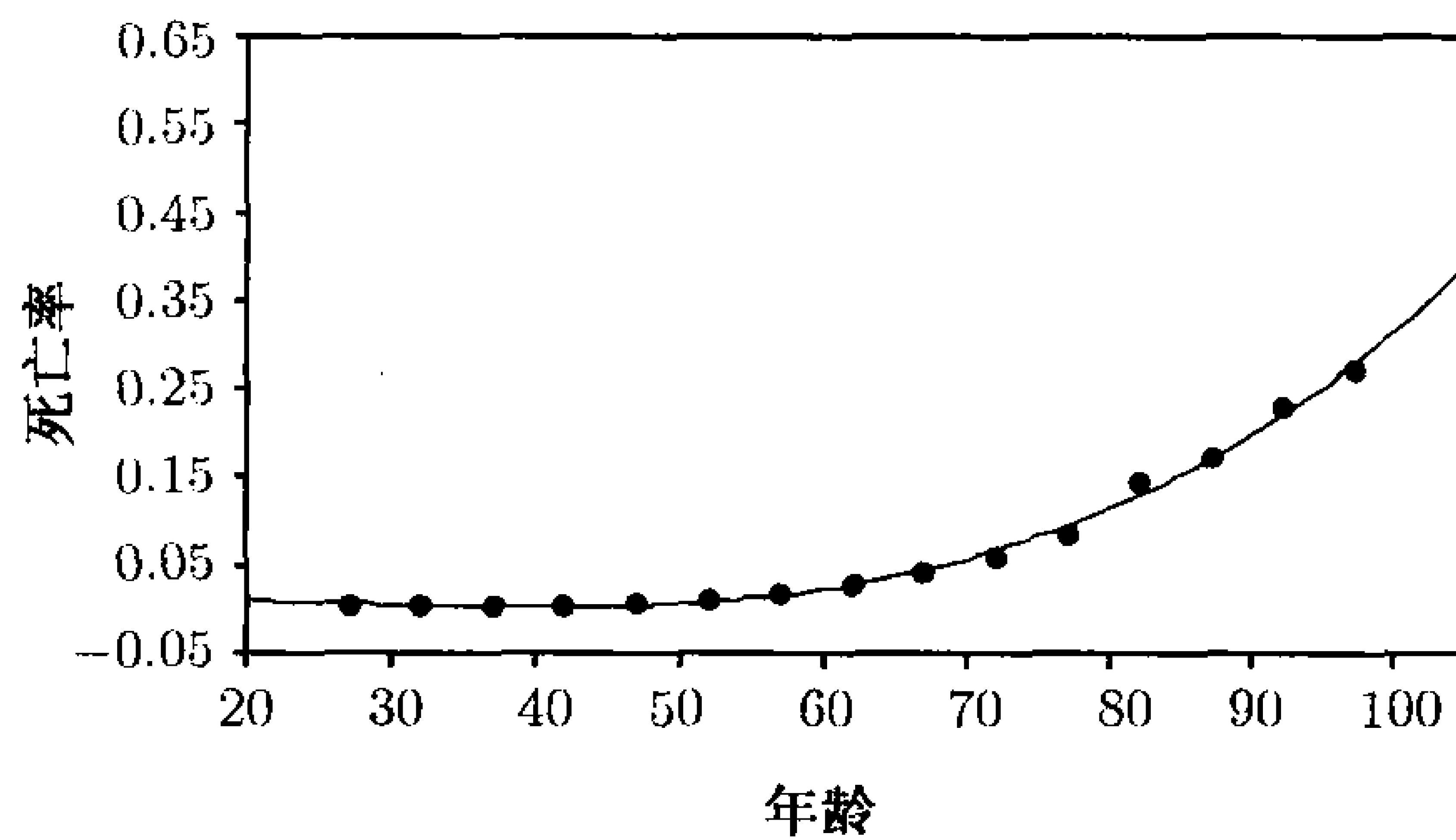


图 15-4 三阶多项式拟合

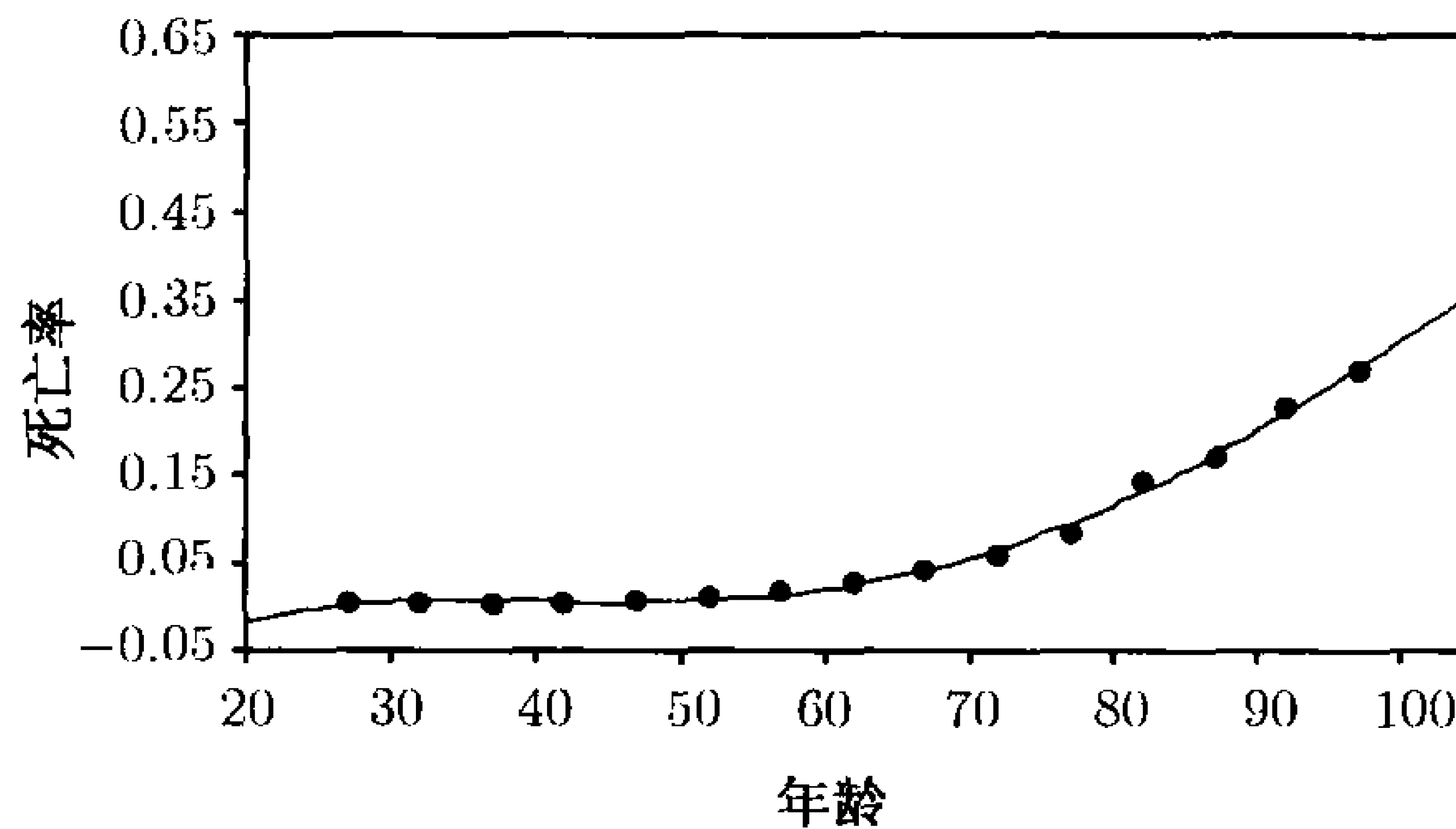


图 15-5 四阶多项式拟合

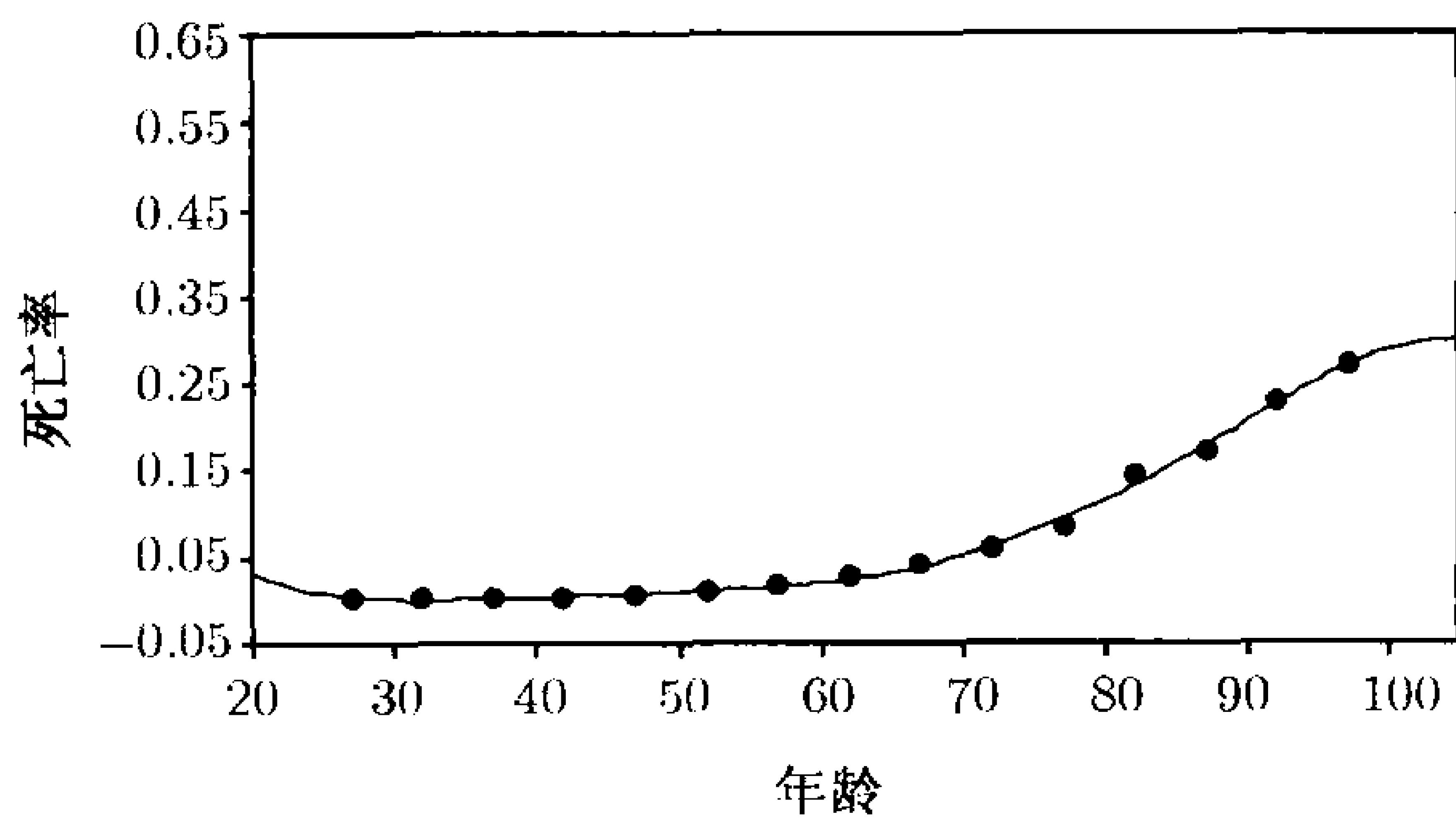


图 15-6　五阶多项式拟合

项式拟合. 需要注意的是, 随着阶数的增加, 拟合的情况也越来越好. 这是因为阶数的增加带来了自由度的增加. 随着阶数的增加, 只有低于 27 岁及高于 97 岁的一些年份的拟合值出现了显著的差异. 平滑样条为这种困境提供了一种解决方法. 事实上, 平滑样条与插值样条类似, 只是并不要求平滑样条必须严格穿过所有的点, 而是要求尽量接近数据点. 三次样条将限制多项式的阶数为 3.

习题

15.1　试对点 (2, 50), (4, 25), (5, 20) 进行多项式插值.

15.2　试给出在最小均方标准下拟合习题 15.1 数据的直线方程.

15.3　三次样条插值

三次样条由分段的三次方程组连接而成, 其特点是在任何点的一阶与二阶导数都是连续的, 而不会出现直接用多项式组缝合时在连接点不可导的情况.

三次样条方法广泛应用于计算机辅助设计, 以及制造业中. 用它来创造触觉与视觉上都显得平滑的表面. 称用三次样条拟合的一系列点为**节点**, 这些节点给出了设计目标或制造目标的基本形状.

在 19 世纪早期由精算师创造的修匀理论的术语中, 称三次样条插值为**接吻插值**①.

在很多数学与工程软件中都嵌有专门的三次样条程序, 所以很容易进行应用.

定义 15.2　令 $\{(x_j, y_j); j=0,\cdots,n\}$ 为 $n+1$ 个不同的节点, 且 $x_0<x_1<x_2<\cdots<x_n$. 如果存在着 n 个系数分别为 a_j, b_j, c_j 和 d_j 的三次多项式 $f_j(x)$, 且满足下列性质

① 之所以采用 "接吻动作" 这种称呼, 是因为连续的三次多项式在节点附近的连续性吻合表现正如接吻一样严密.

Ⅰ. $f(x)=f_j(x)=a_j+b_j(x-x_j)+c_j(x-x_j)^2+d_j(x-x_j)^3$,
对$x_j \leqslant x \leqslant x_{j+1}$ 且 $j=0,1,\cdots,n-1$.
Ⅱ. $f(x_j)=y_j, j=0,1,\cdots,n$.
Ⅲ. $f_j(x_{j+1})=f_{j+1}(x_{j+1}), j=0,1,2,\cdots,n-2$.
Ⅳ. $f_j'(x_{j+1})=f_{j+1}'(x_{j+1}), j=0,1,2,\cdots,n-2$.
Ⅴ. $f_j''(x_{j+1})=f_{j+1}''(x_{j+1}), j=0,1,2,\cdots,n-2$.

则称 $f(x)$ 为一个三次样条.

性质Ⅰ表明 $f(x)$ 由三次多项式拼接而成. 性质Ⅱ表明这个三次多项式组穿过了所有给定的点. 性质Ⅲ要求样条在点集的所有内点处保持连续, 而性质Ⅳ和Ⅴ通过要求一阶导与二阶导的连续性保证了在内点上的平滑.

三次样条的构造

由于每个三次多项式都有 4 个未知系数: a_j, b_j, c_j, d_j. 而在三次样条中要用到 n 个这样的多项式, 因此共需求出 $4n$ 个系数. 性质Ⅱ ~ Ⅴ分别提供了 $n+1, n-1, n-1$ 和 $n-1$ 个方程, 共计有 $4n-2$ 个方程. 为了解出 $4n$ 个系数还需要另外 2 个方程. 一种解决方法是在 x_0 和 x_n 处增加两个与 $f'(x), f''(x)$ 或 $f'''(x)$ 有关的**端点限制函数**. 对端点限制的不同选择会带来不同的结果. 我们将在 15.4 节讨论各种可行的端点限制.

为了构造三次部分, 首先考虑二阶导 $f_j''(x)$, 由于 $f_j(x)$ 是三次的, 所以它是一个线性方程. 因此, 二阶导的拉格朗日表示为

$$f_j''(x)=f''(x_j)\frac{x-x_{j+1}}{x_j-x_{j+1}}+f''(x_{j+1})\frac{x-x_j}{x_{j+1}-x_j}. \tag{15.4}$$

为了简化, 令 $m_j=f''(x_j)$ 和 $h_j=x_{j+1}-x_j$, 则有

$$f_j''(x)=\frac{m_j}{h_j}(x_{j+1}-x)+\frac{m_{j+1}}{h_j}(x-x_j). \tag{15.5}$$

其中 $x_j \leqslant x \leqslant x_{j+1}$ 且 $j=0,1,\cdots,n-1$.

将上式积分两次, 得到

$$f_j(x)=\frac{m_j}{6h_j}(x_{j+1}-x)^3+\frac{m_{j+1}}{6h_j}(x-x_j)^3+p_j(x_{j+1}-x)+q_j(x-x_j), \tag{15.6}$$

这里的 p_i 和 q_j 是积分产生的未知常数. 为检验上式的正确性, 只要将 (15.6) 式求二阶导即可.

由于 $f_j(x_j)=y_j, f_j(x_{j+1})=y_{j+1}$, 代入 (15.6) 式, 可得

$$y_j=\frac{m_j}{6}h_j^2+p_jh_j, \tag{15.7}$$

$$y_{j+1} = \frac{m_{j+1}}{6} h_j^2 + q_j h_j. \tag{15.8}$$

由 (15.7) 式和 (15.8) 式, 可以得到 p_i 和 q_j. 将它们的表达式再代回到 (15.6) 式, 得

$$\begin{aligned} f_j(x) =& \frac{m_j}{6h_j}(x_{j+1}-x)^3 + \frac{m_{j+1}}{6h_j}(x-x_j)^3 + \left(\frac{y_j}{h_j} - \frac{m_j h_j}{6}\right)(x_{j+1}-x) \\ &+ \left(\frac{y_{j+1}}{h_j} - \frac{m_{j+1}h_j}{6}\right)(x-x_j). \end{aligned} \tag{15.9}$$

注意此时 $m_j = f''(x_j)$ 仍然未知. 对 (15.9) 式求导, 得

$$f_j'(x) = -\frac{m_j}{2h_j}(x_{j+1}-x)^2 + \frac{m_{j+1}}{2h_j}(x-x_j)^2 - \left(\frac{y_j}{h_j} - \frac{m_j h_j}{6}\right) + \frac{y_{j+1}}{h_j} - \frac{m_{j+1}h_j}{6}. \tag{15.10}$$

将 $x = x_j$ 带入上式, 化简可得

$$f_j'(x_j) = -\frac{m_j}{3}h_j - \frac{m_{j+1}}{6}h_j + \frac{y_{j+1}-y_j}{h_j}. \tag{15.11}$$

将 (15.10) 式的 j 用 $j-1$ 代替, 再令 $x = x_j$, 可得

$$f_{j-1}'(x_j) = \frac{m_j}{3}h_{j-1} + \frac{m_{j-1}}{6}h_{j-1} + \frac{y_j - y_{j-1}}{h_{j-1}}. \tag{15.12}$$

现在考虑, 性质Ⅳ强制要求在节点的导数相同. 也就是要求 (15.11) 式和 (15.12) 式的右边相同. 由此可以推出 m_{j-1}, m_j 与 m_{j+1} 之间的关系

$$h_{j-1}m_{j-1} + 2(h_{j-1}+h_j)m_j + h_j m_{j+1} = 6\left(\frac{y_{j+1}-y_j}{h_j} - \frac{y_j - y_{j-1}}{h_{j-1}}\right), \tag{15.13}$$

其中 $j = 1, 2, \cdots, n-1$.

(15.13) 式实际上由 $n-1$ 个方程组成, 其中有 $n+1$ 个未知数 $m_0, m_1, \cdots, m_n$. 为了确定 m_0 和 m_n 的值, 需要增加 2 个端点约束方程. 这样就可以通过 (15.13) 式解出剩下的 $n-1$ 个未知数. 这样其实也就得到了 (15.9) 式的全部的解, 也就得到了整个的三次样条.

为了简化表达式, 将 (15.13) 式改写为

$$h_{j-1}m_{j-1} + g_j m_j + h_j m_{j+1} = u_j, \quad j = 1, 2, \cdots, n-1, \tag{15.14}$$

其中

$$u_j = 6\left(\frac{y_{j+1}-y_j}{h_j} - \frac{y_j - y_{j-1}}{h_{j-1}}\right), \quad g_j = 2(h_{j-1}+h_j). \tag{15.15}$$

当端点 m_0 和 m_n 由外生决定时，可以将 (15.14) 式改写为矩阵的形式

$$\underbrace{\begin{bmatrix} g_1 & h_1 & 0 & & \cdots & & 0 \\ h_1 & g_2 & h_2 & 0 & & \cdots & 0 \\ 0 & h_2 & g_3 & h_3 & 0 & \cdots & 0 \\ & 0 & & & & & \vdots \\ & & \ddots & \ddots & & h_{n-3} & 0 \\ & & & & h_{n-3} & g_{n-2} & h_{n-2} \\ 0 & 0 & \cdots & & 0 & h_{n-2} & g_{n-1} \end{bmatrix}}_{(n-1)\times(n-1)} \underbrace{\begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ \\ \vdots \\ m_{n-2} \\ m_{n-1} \end{bmatrix}}_{(n-1)\times 1} = \underbrace{\begin{bmatrix} u_1 - h_0 m_0 \\ u_2 \\ \vdots \\ \\ \vdots \\ u_{n-2} \\ u_{n-1} - h_{n-1} m_n \end{bmatrix}}_{(n-1)\times 1} \quad (15.16)$$

或

$$\boldsymbol{H}\boldsymbol{m} = \boldsymbol{v}, \tag{15.17}$$

其中矩阵 $\boldsymbol{H}$ 是一个可逆的三角矩阵. 因此 (15.17) 式有唯一的解 $\boldsymbol{m} = \boldsymbol{H}^{-1}\boldsymbol{v}$. 另外可通过高斯消除法人工求解.

一旦得到了 $m_1, m_2, \cdots, m_{n-1}$ 的值, 就可以通过下式求出 c_j 的值

$$c_j = \frac{m_j}{2}, j = 1, \cdots, n-1.$$

性质 II 说明 $a_j = y_j, j = 0, \cdots, n-1.$

性质 V 说明 $m_j + 6d_j h_j = m_{j+1}, j = 0, \cdots, n-2.$

变形, 可得 $d_j = \dfrac{m_{j+1} - m_j}{6h_j}, j = 0, \cdots, n-2.$

性质III说明 $a_j + b_j h_j + c_j h_j^2 + d_j h_j^3 = y_{j+1}, j = 0, \cdots, n-2.$

将上面所得到的 a_j, c_j, d_j 的表达式带入, 可得:

$$b_j = \frac{y_{j+1} - y_j}{h_j} - \frac{h_j(2m_j + m_{j+1})}{6}, \quad j = 0, \cdots, n-2.$$

总之, 可以解出样条前 $n-1$ 个部分的系数为

$$a_j = y_j, \quad b_j = \frac{y_{j+1} - y_j}{h_j} - \frac{h_j(2m_j + m_{j+1})}{6},$$

$$c_j = \frac{m_j}{2}, \quad d_j = \frac{m_{j+1} - m_j}{6h_j}, \quad j = 0, \cdots, n-2. \tag{15.18}$$

这样现存的唯一问题就是对端点约束函数的选择. 当然有很多可能的选择, 一旦做出了一个选择, 也就完全确定了这 n 个三次方程. 因此通过 (15.18) 式就可以得到 b_{n-1}, c_{n-1} 和 d_{n-1} 的值.

情形 1：简单三次样条　($m_0 = m_n = 0$)

简单样条的做法是设 (15.16) 式的 m_0 和 m_n 为零. 因为 m_0 和 m_n 是在端点的二阶导数, 设其为零则将两个端点的波动控制到最小. 它同样保证了样条在两个节点端点之外的线性, 因为它使得在数据两个端点之外的波动达到了最小. 在许多考虑数据点外推的情况下, 这通常是最安全的假设. 但应注意的是二阶导数的端点约束本身并没有对端点处的斜率有所限制.

情形 2：曲率调整的三次样条　(m_0 和 m_n 为固定值)

将端点处的二阶导数 m_0 和 m_n 的值固定为预先给定的 $f''(x_0)$ 和 $f''(x_n)$ 的做法与情形 1 相类似. 这时同样可以直接用 (15.16) 式得到 $m_1, m_2, \cdots, m_{n-1}$ 的值. 然而, 在实际中很难不考虑主观的判断. 一般的做法都是从简单三次样条开始. 如果需要对尾部有更高的曲率要求, 就可以用情形 2 进行调整.

其他的端点约束方法将更加复杂, 有些可能需要对 (15.14) 式的第一个与最后一个方程做一些调整, 这也就意味着对 (15.17) 式中的矩阵 $\boldsymbol{H}$ 和向量 $\boldsymbol{v}$ 进行调整.

情形 3：抛物线偏转样条　($m_0 = m_1, m_n = m_{n-1}$)

这个方法将第一个与最后一个区间的两个三次方程降为二次方程, 也就是增加了两个新的约束：$d_0 = 0$ 和 $d_n = 0$. 这样也就保证在首尾两个区间的二阶导数是相等的. 即：$m_0 = m_1, m_n = m_{n-1}$. 因此, (15.14) 式的首尾两个方程也要改写为

$$(3h_0 + 2h_1)m_1 + h_1 m_2 = u_1,$$

$$h_{n-2}m_{n-2} + (2_{n-2} + 3h_{n-1})m_{n-1} = u_{n-1.} \tag{15.19}$$

情形 4：三次偏转样条

这种方法要求在 $[x_0,x_1]$ 区间的三次方程为 $[x_1, x_2]$ 区间内方程的拓展. 也就是说, 在整个 $[x_0, x_2]$ 区间采用同一个三次函数. 这也是所谓的无节点条件. 类似地对另一端同时采取相同的做法.

可以要求端点的三阶导数分别与 x_1 和 x_{n-1} 相等来满足假设. 即

$$f_0'''(x_1) = f_1'''(x_1), \quad f_{n-2}'''(x_{n-1}) = f_{n-1}'''(x_{n-1}).$$

因为在 $[x_0, x_2]$ 与 $[x_{n-2}, x_n]$ 中样条的三阶导都是常数, 则这两个区间内的二阶导数为一个线性方程. 因此, 在 $[x_0, x_2]$ 与 $[x_{n-2}, x_n]$ 内部的任何子区间中的二阶导数的斜率都是相同的. 用方程可表示为

$$\frac{m_1 - m_0}{h_0} = \frac{m_2 - m_1}{h_1},$$

$$\frac{m_n - m_{n-1}}{h_{n-1}} = \frac{m_{n-1} - m_{n-2}}{h_{n-2}},$$

或者等价地, 有

$$m_0 = m_1 - \frac{h_0(m_2 - m_1)}{h_1},$$

$$m_n = m_{n-1} + \frac{h_{n-1}(m_{n-1} - m_{n-2})}{h_{n-2}}. \tag{15.20}$$

这样, (15.14) 式的首尾方程被替换为

$$\left(3h_0 + 2h_1 + \frac{h_0^2}{h_1}\right) m_1 + \left(h_1 - \frac{h_0^2}{h_1}\right) m_2 = u_1,$$

$$\left(h_{n-2} - \frac{h_{n-1}^2}{h_{n-2}}\right) m_{n-2} + \left(2h_{n-2} + 3h_{n-1} + \frac{h_{n-1}^w}{h_{n-2}}\right) m_{n-1} = u_{n-1}. \tag{15.21}$$

情形 5: 强制性三次样条

这种方法固定在两个端点的斜率值为 $f_0'(x_0)$ 和 $f_{n-1}'(x_n)$. 这时由 (15.11) 式和 (15.12) 式, 二阶导数为

$$m_0 = \frac{3}{h_0}\left(\frac{y_1 - y_0}{h_0} - f_0'(x_0)\right) - \frac{m_1}{2},$$

$$m_n = \frac{3}{h_{n-1}}\left(f_{n-1}'(x_n) - \frac{y_n - y_{n-1}}{h_{n-1}}\right) - \frac{m_{n-1}}{2}. \tag{15.22}$$

所以 (15.14) 的首尾两个方程被替换为

$$\left(\frac{3}{2}h_0 + 2h_1\right) m_1 + h_1 m_2 = u_1 - 3\left(\frac{y_1 - y_0}{h_0} - f_0'(x_0)\right),$$

$$h_{n-2}m_{n-2} + \left(2h_{n-2} + \frac{3}{2}h_{n-1}\right) m_{n-1} = u_{n-1} - 3\left(f_{n-1}'(x_n) - \frac{y_n - y_{n-1}}{h_{n-1}}\right).$$

例 15.3 利用条件 I ~ V 强制求解通过点 (2, 50), (4, 25) 以及 (5, 20) 的三次样条, 这时的强制边界条件为 $f'(2) = -25, f'(5) = -4$.

解 设在区间 $x_0 = 2$ 到 $x_1 = 4$ 之间的三次样条为多项式

$$f_0(x) = 50 + b_0(x-2) + c_0(x-2)^2 + d_0(x-2)^3.$$

同样地, 令在 $x_1 = 4$ 到 $x_2 = 5$ 区间内的样条为多项式

$$f_1(x) = 25 + b_1(x-4) + c_1(x-4)^2 + d_1(x-4)^3.$$

因此需要解出 $b_0, c_0, d_0, b_1, c_1, d_1$ 这 6 个未知数. 首先由 $f_0(x)$ 与 $f_1(x)$ 的定义可以得到

$$f_0(4) = 50 + 2b_0 + 4c_0 + 8d_0 = 25,$$

$$f_1(5) = 25 + b_1 + c_1 + d_1 = 20.$$

由 $x = 4$ 的平滑性质, 可以得到

$$f_0'(4) = b_0 + 2c_0(4-2) + 3d_0(4-2)^2 = f_1'(4) = b_1,$$

$$f_0''(4) = 2c_0 + 6d_0(4-2) = f_1''(4) = 2c_1.$$

最后, 由边界条件有

$$f_0'(2) = b_0 = -25, \quad f_1'(5) = b_1 + 2c_1 + 3d_1 = -4.$$

因此, 我们得到含 6 个未知数的 6 个线性方程. 用矩阵可表示为

$$\begin{bmatrix} 2 & 4 & 8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 4 & 12 & -1 & 0 & 0 \\ 0 & 2 & 12 & 0 & -2 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} b_0 \\ c_0 \\ d_0 \\ b_1 \\ c_1 \\ d_1 \end{bmatrix} = \begin{bmatrix} -25 \\ -5 \\ 0 \\ 0 \\ -25 \\ -4 \end{bmatrix}$$

可通过逐步消除法求解方程组. 首先得到 $b_0 = -25$, 则可以将矩阵化简为

$$\begin{bmatrix} 4 & 8 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 4 & 12 & -1 & 0 & 0 \\ 2 & 12 & 0 & -2 & 0 \\ 0 & 0 & 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} c_0 \\ d_0 \\ b_1 \\ c_1 \\ d_1 \end{bmatrix} = \begin{bmatrix} 25 \\ -5 \\ 25 \\ 0 \\ -4 \end{bmatrix}$$

由 $c_0 = 6.25 - 2d_0$, 化简得

$$\begin{bmatrix} 0 & 1 & 1 & 1 \\ 4 & -1 & 0 & 0 \\ 8 & 0 & -2 & 0 \\ 0 & 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} d_0 \\ b_1 \\ c_1 \\ d_1 \end{bmatrix} = \begin{bmatrix} -5 \\ 0 \\ -12.5 \\ -4 \end{bmatrix}$$

由 $d_0 = 0.25b_1$, 可得

$$\begin{bmatrix} 1 & 1 & 1 \\ 2 & -2 & 0 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} b_1 \\ c_1 \\ d_1 \end{bmatrix} = \begin{bmatrix} -5 \\ -12.5 \\ -4 \end{bmatrix}$$

由 $b_1 = -6.25 + c_1$, 得到

$$\begin{bmatrix} 2 & 1 \\ 3 & 3 \end{bmatrix} \begin{bmatrix} c_1 \\ d_1 \end{bmatrix} = \begin{bmatrix} 1.25 \\ 2.25 \end{bmatrix}$$

最后, 由 $c_1 = 0.625 - 0.5d_1$ 可以解出 $d_1 = 0.25$. 逆推回去, 可以得到最终解

$$b_0 = -25, \quad c_0 = 9.125, \quad d_0 = -1.4375, \quad b_1 = -5.75, \quad c_1 = 0.5, \quad d_1 = 0.25.$$

这样, 我们就得到了最终的三次插值样条

$$f(x) = \begin{cases} 50 - 25(x-2) + 9.125(x-2)^2 - 1.4375(x-2)^3, & 2 \leqslant x \leqslant 4, \\ 25 - 5.75(x-4) + 0.5(x-4)^2 + 0.25(x-4)^3, & 4 \leqslant x \leqslant 5. \end{cases}$$

图 15-7 展示了例 15.3 中的插值三次样条以及对应的简单三次样条. 图中还给出了函数 $h(x) = 100/x$ 的曲线, 该函数同样穿过了这 3 个节点. 强制样条在端点处的斜率与函数 $h(x)$ 的斜率相同 (由例题中的强制边界条件确定). 这样就使得强制样条的曲线比简单样条明显地更加接近 $h(x)$. 由于简单样条要求在端点处的二阶导为零, 因此使得它在两个端点看起来更接近一条直线. □

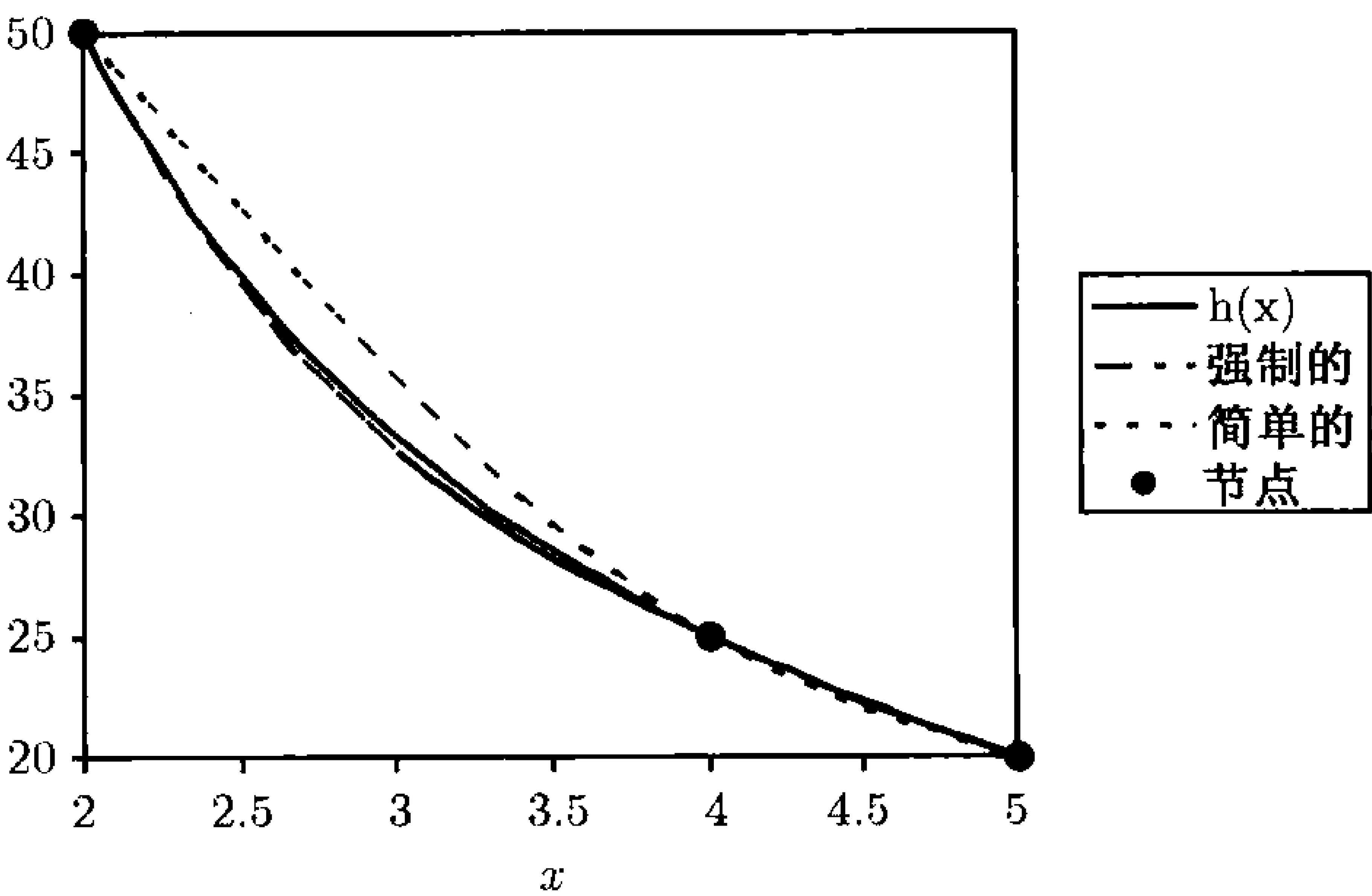


图 15-7 例 15.3 的强制样条与简单样条

本节的三次样条均穿过全部的节点. 这个限制条件在平滑的要求下将会改变. 平滑样条将在 15.6 节中介绍.

例 15.4 表 15-1 的最后一列数据是 15 个 5 年区间内的年死亡率. 我们用 95 ~ 99 来表示最后一个区间. 采用一个简单三次样条对这些数据进行插值, 并将区间内的死亡率看作是每个区间中点年龄的年死亡率, 并将这些中点年龄看作简单三次样条的节点. 图 15-8 给出了依对数死亡率拟合出来的插值三次样条. 样条公式由定义 15.2 的性质 I 给出. 样条的所有系数在表 15-2 中给出.

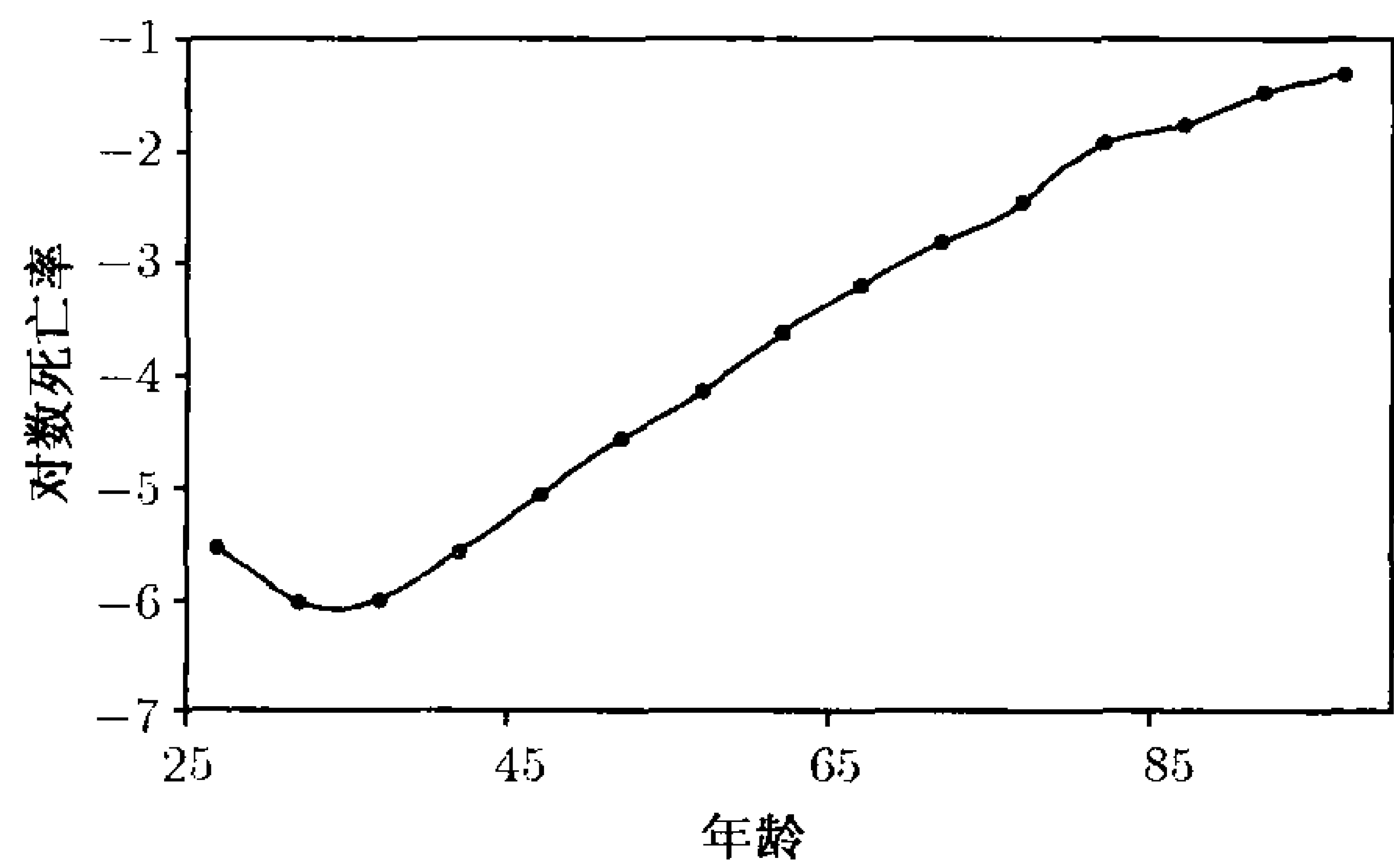


图 15-8 例 15.4 中死亡率数据的三次样条拟合

表 15-2 例 15.4 的样条系数

j	x_j	a_j	b_j	c_j	d_j
0	27	3.8936×10^{-3}	-3.5093×10^{-4}	0	2.5230×10^{-6}
1	32	2.4543×10^{-3}	-1.6171×10^{-4}	3.7844×10^{-5}	-8.4886×10^{-7}
2	37	2.4857×10^{-3}	1.5307×10^{-4}	2.5112×10^{-5}	-5.1079×10^{-7}
3	42	3.8150×10^{-3}	3.6587×10^{-4}	1.7450×10^{-5}	2.0794×10^{-6}
4	47	6.3405×10^{-3}	6.9632×10^{-4}	4.8640×10^{-5}	-4.3460×10^{-6}
5	52	1.0495×10^{-2}	8.5678×10^{-4}	-1.6550×10^{-5}	1.2566×10^{-5}
6	57	1.5936×10^{-2}	1.6337×10^{-3}	1.7194×10^{-4}	-1.1922×10^{-5}
7	62	2.6913×10^{-2}	2.4590×10^{-3}	-6.8828×10^{-6}	1.3664×10^{-5}
8	67	4.0744×10^{-2}	3.4150×10^{-3}	1.9808×10^{-4}	-2.5761×10^{-5}
9	72	5.9551×10^{-2}	3.4638×10^{-3}	-1.8833×10^{-4}	1.1085×10^{-4}
10	77	8.6018×10^{-2}	9.8939×10^{-3}	1.4744×10^{-3}	-2.1542×10^{-4}
11	82	1.4542×10^{-1}	8.4813×10^{-3}	-1.7569×10^{-3}	2.2597×10^{-4}
12	87	1.7215×10^{-1}	7.8602×10^{-3}	1.6327×10^{-3}	-1.7174×10^{-4}
13	92	2.3080×10^{-1}	1.1306×10^{-2}	-9.4349×10^{-4}	6.2899×10^{-5}

习题

15.3 若将例 15.3 中的强制边界条件去掉, 求简单三次样条.

15.4 对点 (−2, 0), (−1, 1), (0, 0), (1, 1) 和 (2, 0), 用 (15.16) 式的矩阵方程求解简单三次样条.

15.5 判断下面的方程是否为三次样条函数.

(a)

$$f(x)=\begin{cases} x, & -4\leqslant x\leqslant 0,\\ x^3+x, & 0\leqslant x\leqslant 1,\\ 3x^2-2x+1, & 1\leqslant x\leqslant 9.\end{cases}$$

(b)

$$f(x)=\begin{cases} x^3, & 0\leqslant x\leqslant 1,\\ 3x^2-3x+1, & 1\leqslant x\leqslant 2,\\ x^3-4x^2+13x-11, & 2\leqslant x\leqslant 4.\end{cases}$$

(c)

$$f(x)=\begin{cases} x^3+2x, & -1\leqslant x\leqslant 0,\\ 2x^2+2x, & 0\leqslant x\leqslant 1,\\ x^3-x^2+5x-1, & 1\leqslant x\leqslant 3.\end{cases}$$

15.6 求出 a,b,c 的值满足以下三次样条函数.

$$f(x)=\begin{cases} x^3+4, & 0\leqslant x\leqslant 1,\\ a+b(x-1)+c(x-1)^2+4(x-1)^3, & 1\leqslant x\leqslant 3.\end{cases}$$

15.7 求出在 $x=-1,0,1$ 点与 $\sin(x\pi/2)$ 取值相同的强制三次样条.

15.8 对于下面的函数:

$$f(x)=\begin{cases} 28+25x+9x^2+x^3, & -3\leqslant x\leqslant -1,\\ 26+19x+3x^2-x^3, & -1\leqslant x\leqslant 0,\\ 26+19x+3x^2-2x^3, & 0\leqslant x\leqslant 3,\\ -163+208x-60x^2+5x^3, & 3\leqslant x\leqslant 4.\end{cases}$$

(a) 证明 $f(x)$ 是三次样条函数.

(b) 判断 5 个端点条件中, 哪些条件已经在样条判断中使用过.

15.4 样条近似函数

当样条作为某个连续函数的近似时, 简单样条与强制三次样条都具有特别令人满意的性质. 例如, 考虑函数

$$h(x)=\frac{100}{x},\quad 2\leqslant x\leqslant 5.$$

该函数在节点 $x=2,4,5$ 的取值与例 15.3 相同. 假如认为这些节点确实来自函数 $h(x)$, 这样, 就可以认为插值三次样条是函数 $h(x)$ 的一个近似. 在很多应用中, 例如计算机作图中需要平滑图像, 这些平滑的图像就可以通过选择有限数量的节点以及一个三次样条插值系统进行有效地表示.

平滑性可以通过整个函数的曲率度量, 最常用的度量方法是平方准则

$$S=\int_{x_0}^{x_n}[f''(x)]^2\,\mathrm{d}x, \tag{15.23}$$

它代表全部的二阶导数的平方.

现在, 考虑任何一个连续函数 $h(x)$, 它在区间 $[x_0,x_n]$ 的一阶导与二阶导同样是连续的. 假设已选择了 $n-1$ 个内部节点 $\{x_j,h(x_j)\}_{j=1}^{n-1}$, $x_0<x_1<x_2<\cdots<x_n$. 令 $f(x)$ 为一个穿过上面所有节点的三次样条, 同时在端点处满足下面的条件之一

$$f'(x_0)=h'(x_0)\ \text{且}\ f'(x_n)=h'(x_n)\quad\text{(强制样条)}$$

或

$$f''(x_0) = 0 \text{ 且 } f''(x_n) = 0. \quad \text{(简单样条)}$$

这种简单或者强制样条比其他任何只是通过这 $n+1$ 个节点的函数 $h(x)$ 而言, 具有最小的总曲率, 详见下面的定理.

定理 15.5 令 $f(x)$ 是穿过 $n+1$ 个给定节点的简单或强制三次样条. 令 $h(x)$ 是任何具有连续一阶导与二阶导的同样穿过这些节点的函数. 同时, 对于强制样条有 $f'(x_0) = h'(x_0)$ 且 $f'(x_n) = h'(x_n)$. 则

$$\int_{x_0}^{x_n} [f''(x)]^2 \mathrm{d}x \leqslant \int_{x_0}^{x_n} [h''(x)]^2 \mathrm{d}x. \tag{15.24}$$

证明 令 $D(x) = h(x) - f(x)$, 则 $D''(x) = h''(x) - f''(x)$, 可得

$$[h''(x)]^2 = [f''(x)]^2 + [D''(x)]^2 + 2f''(x)D''(x).$$

两边积分, 得到

$$\int_{x_0}^{x_n} [h''(x)]^2 \mathrm{d}x = \int_{x_0}^{x_n} [f''(x)]^2 \mathrm{d}x + \int_{x_0}^{x_n} [D''(x)]^2 \mathrm{d}x + 2\int_{x_0}^{x_n} f''(x)D''(x)\mathrm{d}x.$$

只要能够证明 $\int_{x_0}^{x_n} f''(x)D''(x)\mathrm{d}x = 0$, 则函数 $h(x)$ 的总曲率 $\int_{x_0}^{x_n} [h''(x)]^2 \mathrm{d}x$ 就等于样条的总曲率 $\int_{x_0}^{x_n} [f''(x)]^2 \mathrm{d}x$ 加上一个非负的量 $\int_{x_0}^{x_n} [D''(x)]^2 \mathrm{d}x$, 进而定理也就得证.

通过分部积分, 得到

$$\int_{x_0}^{x_n} f''(x)D''(x)\mathrm{d}x = f''(x)D'(x)\bigg|_{x_0}^{x_n} - \int_{x_0}^{x_n} f'''(x)D'(x)\mathrm{d}x.$$

对于强制三次样条, 边界条件意味着

$$D'(x_0) = h'(x_0) - f'(x_0) = 0, \quad D'(x_n) = h'(x_n) - f'(x_n) = 0.$$

所以分部积分式的第一项为零.

对于简单三次样条, 由于 $f''(x_0) = f''(x_n) = 0$, 同样分部积分的第一项为零. 而分部积分式的第二项可以分解为

$$\int_{x_0}^{x_n} f'''(x)D'(x)\mathrm{d}x = \sum_{j=0}^{n-1} \int_{x_j}^{x_{j+1}} f'''(x)D'(x)\mathrm{d}x.$$

再对每一个区间分部积分, 可得

$$\int_{x_j}^{x_{j+1}} f'''(x)D'(x)\mathrm{d}x = f'''(x)D(x)\bigg|_{x_j}^{x_{j+1}} - \int_{x_j}^{x_{j+1}} f^{(4)}(x)D(x)\mathrm{d}x.$$

由插值条件

$$D(x_j) = h(x_j) - f(x_j) = 0, \quad j = 0, 1, \cdots, n.$$

所以上述分部积分的第一项为零. 也就是说, 我们只考虑了穿过节点的函数 $h(x)$. 上述分部积分的第二项同样为零, 因为 $f(x)$ 是一个三次多项式, 所以四阶导数必然为零. 因此, 对于简单样条和强制样条都有

$$\int_{x_0}^{x_n} f''(x)D''(x)\mathrm{d}x = 0. \qquad \square$$

因此, 如果想要对一系列的数值进行平滑处理, 而且已经对区间两端的斜率有些了解, 那么强制三次样条是最有吸引力的选择. 比如在死亡表的构造中, 在出生的最初几天以及几周内, 由于先天性疾病或者其他对于出生婴儿死亡的影响, 死亡力或风险率有很明显的下降趋势. 在高年龄阶段, 死亡力趋于平缓, 100 岁之上的死亡力维持在 0.3 至 0.4 之间. 采用强制三次样条来拟合观测的死亡力可以得到在各年龄观测区间的尾部满足一定要求的最平滑的函数. 如果死亡力数据仅仅限制在某些年龄区间内 (如寿险与年金险中常常使用的), 则可以在简单三次样条和强制三次样条中任选一个. 增加一个强制条件将增加端点的斜率控制.

例 15.6 对于例 15.3 中得到的强制三次样条, 计算平方准则的曲率值. 对于同样通过给定节点的 $h(x) = 100/x$ 也计算其曲率值.

解 样条函数为

$$f(x) = \begin{cases} 50 - 25(x-2) + 9.125(x-2)^2 - 1.4375(x-2)^3, & 2 \leqslant x \leqslant 4, \\ 25 - 5.75(x-4) + 0.5(x-4)^2 + 0.25(x-4)^3, & 4 \leqslant x \leqslant 5. \end{cases}$$

则其二阶导为

$$f''(x) = \begin{cases} 18.25 - 8.625(x-2) = 35.5 - 8.625x, & 2 \leqslant x \leqslant 4, \\ 1 + 1.5(x-4) = 1.5x - 5, & 4 \leqslant x \leqslant 5. \end{cases}$$

则可以计算总的曲率值为

$$\begin{aligned} \int_2^5 [f''(x)]^2\mathrm{d}x &= \int_2^4 (35.5 - 8.625x)^2\mathrm{d}x + \int_4^5 (1.5x-5)^2\mathrm{d}x \\ &= \int_1^{18.25} y^2 \frac{1}{8.625}\mathrm{d}y + \int_1^{2.5} y^2 \frac{1}{1.5}\mathrm{d}y \\ &= \left.\frac{y^3}{25.875}\right|_1^{18.25} + \left.\frac{y^3}{4.5}\right|_1^{2.5} = 238.125. \end{aligned}$$

$h(x)$ 的二阶导为 $h''(x) = 200x^{-3}$, 因此, 曲率为

$$\int_2^5 (200x^{-3})^2\mathrm{d}x = \int_2^5 40\ 000x^{-6}\mathrm{d}x = -8\ 000x^{-5}\Big|_2^5 = 247.44.$$

注意到函数 $h(x)$ 与强制样条的总曲率相当地接近. 从图 15-7 也可以看出, 两个函数在形状上非常类似, 因此曲率自然也比较接近. 当然, 由定理 15.5 知, 样条的曲率会更小一些, 尽管在本例中二者曲率的差别只是微乎其微. 在习题 15.9 中要求计算相对的简单样条的总曲率 (图 15-7 中同样有显示). 因为它要更 "直" 一些, 所以我们可以预见它的总曲率会明显的小一些, 通过习题 15.9 可以确认这一判断. □

习题

15.9 对于习题 15.3 中得到的简单三次样条, 计算平方准则的曲率值.

15.5 样条的外推

在很多实际问题中我们往往希望得到这样一个模型, 一方面它能够忠实的反应历史数据, 另一方面, 它也能用来预测未来. 例如, 在确定保险公司的负债时, 由于未来索赔的支付与通货膨胀率有关, 所以精算师要对未来 5 年到 10 年的通货膨胀率进行估计. 其中的一种方法就是对通货膨胀率的历史数据用三次样条函数拟合.

简单地将三次样条函数外推到大于 x_n 的区间内可能会使估计值不稳定的波动. 有些情况得到的估计值也相当不合理. 因此, 有必要要求估计值服从比较简单的形式. 特别地, 线性预测在大多数实际情况下都是合理的. 而三次样条函数也可以解决这个问题.

简单三次样条要求在端点处的二阶导为零, 因此自然的外推方法就是从端点依这个斜率延长直线. 当然, 对于任何类型的样条, 都可以采用端点的斜率值进行线性外推. 然而, 除非像简单样条那样二阶导为零, 否则端点处的二阶导条件会比较麻烦. 我们得到在两个端点的外推函数为

$$f(x) = f(x_n) + f'(x_n)(x - x_n), \qquad x > x_n,$$
$$f(x) = f(x_0) - f'(x_0)(x_0 - x), \qquad x < x_0.$$

例 15.7 对例 15.3 的强制样条求外推公式, 并给出在 $x = 0$ 和 $x = 7$ 的外推值.

解 在第一个区间内 $f(x) = 50 - 25(x-2) + 9.125(x-2)^2 - 1.4375(x-2)^3$, 因此 $f(2) = 50, f'(2) = -25$. 故对 $x < 2$, 外推公式为 $f(x) = 50-(-25)(2-x) = 100-25x$. 在最后一个区间内 $f(x) = 25-5.75(x-4)+0.5(x-4)^2+0.25(x-4)^3$, 因此 $f(5) = 20$, $f'(5) = -4$. 所以对于 $x > 5$, 外推公式为 $f(x) = 20 - 4(x-5) = 40 - 4x$. 在点 $x = 0$, 外推值为 $100 - 25(0) = 100$, 而在点 $x = 7$ 处, 外推值为 $40 - 4(7) = 12$. □

习题

15.10 对例 15.3, 给出其在简单样条下的外推公式, 并给出在 $x = 0$ 和 $x = 7$ 的外推值.

15.6 平滑样条

在很多的精算应用中, 除了对观测数据进行插值, 我们往往会要求更多. 如果数据包含了一个随机 (噪声) 因素, 最好的做法往往是要求三次样条或者其他的平滑函数能够比较靠近数据点, 而不是要求函数必须准确无误地穿过每一个数据点.

在精算师于 20 世纪早期发展起来的修匀理论的术语中, 称平滑样条为**修正的接吻插值**, 其中"修正的" 一词表明在连接点 (节点) 的取值是通过原始数据调整的.

平滑三次样条的技术发展与插值三次样条相同, 只是原始节点的数据 (x_i, y_i) 用调整过的节点 (x_j, a_j) 代替. 纵坐标 a_j 是平滑三次样条方程的常数项

$$f_j(x) = a_j + b_j(x - x_j) + c_j(x - x_j)^2 + d_j(x - x_j)^3. \tag{15.25}$$

首先假设原始数据的纵坐标点满足下面的模型:

$$y_j = g(x_j) + \varepsilon_j,$$

其中 ε_j, $j = 0, 1, \cdots, n$, 是独立分布的随机变量, 均值为 0, 方差为 σ_j^2, $g(x)$ 是某个具有优良性质的函数[①].

例 15.8 *在各年龄 j 的死亡率 q_i 的估计方法为死亡人数与总人数之比 D_j/n_j, 其中 D_j 为二项分布 (n_j, q_j) 的随机变量. 而死亡率的估计值 $\hat{q}_j = d_j/n_j$, 其中 d_j 是观测到的死亡人数, 其方差为 $\sigma_j^2 = q_j(1 - q_j)/n_j$, 可以用 $\hat{q}_j(1 - \hat{q}_j)/n_j$ 近似.*

我们试图找到某个平滑函数 $f(x)$, 这里只考虑三次样条, 可以作为 "真实" 函数 $g(x)$ 的近似. 因为 $g(x)$ 是具有优良性质的, 所以要求 $f(x)$ 本身也尽可能地平滑. 另一方面, 我们也希望尽可能地忠于给定的数据点. 应该注意到这两个要求其实是互相冲突的, 因此需要在拟合与平滑之间做出妥协.

拟合程度可以通过卡方标准来度量

$$F = \sum_{j=0}^{n} \left(\frac{y_j - a_j}{\sigma_j} \right)^2. \tag{15.26}$$

这是一种标准的用来度量拟合程度的统计准则. 我们在 13.4.3 节中已经讨论过, 它是具有 $n + 1$ 个自由度的卡方分布[②].

① 这里不给出所谓 "优良性质" 的定义, 而仅仅说 $g(x)$ 在一般意义下比较平滑. 至少, 我们需要函数的一阶导与二阶导连续.

② 不损失任何的自由度. 与拟合优度检验不同的是, 如果和式中只有一项未知其他都已知时, 仍然无法推断这个值.

平滑程度可以由三次样条的总平滑度度量. 平滑度, 也称作总曲率, 可由下面这个平方正规平滑标准来度量

$$S=\int_{x_0}^{x_n}\left[f''(x)\right]^2\mathrm{d}x,$$

定理 15.5 证明了在诸多一阶导与二阶导连续的函数中, 简单三次样条与强制三次样条都能够使得平方标准达到最小. 这也是对选择三次样条作为平滑函数的支持.

为了更加明确拟合与平滑之间的冲突, 我们建立一个评判标准, 它是对拟合与平滑优度的一种加权平均. 令

$$L=pF+(1-p)S=p\sum_{j=0}^{n}\left(\frac{y_j-a_j}{\sigma_j}\right)^2+(1-p)\int_{x_o}^{x_n}[f''(x)]^2\mathrm{d}x.$$

参数 p 反应了赋予冲突两方的重要程度. 一方面表示是否足够接近数据, 而另一方面表示是否为一个平滑的曲线. 显然线性函数能够满足方程

$$S=\int_{x_0}^{x_n}[f''(x)]^2\mathrm{d}x=0,$$

这也就意味着, 在 $p=0$ 的极限情况下只对平滑性有约束, 此时的样条函数 $f(x)$ 将为直线. 在另一种极端下, 如果 $p=1$, 表示我们只关心样条是否接近观测点集. 这时得到的就是一个插值样条, 它可以准确地穿过所有的点.

因为样条是分段的三次函数, 因此, 平滑性准则可以表示为

$$S=\int_{x_0}^{x_n}[f''(x)]^2\mathrm{d}x=\sum_{j=0}^{n-1}\int_{x_j}^{x_{j+1}}[f_j''(x)]^2\mathrm{d}x.$$

由 (15.5) 式, 有

$$f_j''(x)=\frac{m_j}{h_j}(x_{j+1}-x)+\frac{m_{j+1}}{h_j}(x-x_j),$$

因此可以得到

$$\begin{aligned}\int_{x_j}^{x_{j+1}}[f_j''(x)]^2\mathrm{d}x&=\int_{x_j}^{x_{j+1}}\left[\frac{m_j}{h_j}(x_{j+1}-x)+\frac{m_{j+1}}{h_j}(x-x_j)\right]^2\mathrm{d}x\\&=\int_0^1[m_j(1-y)+m_{j+1}y]^2h_j\mathrm{d}y=h_j\int_0^1[m_j+(m_{j+1}-m_j)y]^2\mathrm{d}y\\&=h_j\left.\frac{[m_j+(m_{j+1}-m_j)y]^3}{3(m_{j+1}-m_j)}\right|_0^1=\frac{h_j}{3}(m_j^2+m_jm_{j+1}+m_{j+1}^2),\end{aligned}$$

注意在第二行中, 用 $y=(x-x_j)/h_j$ 进行替换. 这样, 原判别函数变为

$$L=p\sum_{j=0}^{n}\left(\frac{y_j-a_j}{\sigma_j}\right)^2+(1-p)\sum_{j=0}^{n-1}\frac{h_j}{3}(m_j^2+m_jm_{j+1}+m_{j+1}^2).$$

我们需要通过改变 $2n+2$ 个未知数 $\{a_j, m_j; j=0,\cdots,n\}$ 的值来使得该函数达到最小. 注意到在解出这些变量之后, 每个区间 $[x_j, x_{j+1}]$ 上会有 4 个信息量 $\{a_j, a_{j+1}, m_j, m_{j+1}\}$. 这使得我们能够完全确定每个区间内的三次插值样条. 现在研究如何解出这些变量.

首先考虑简单平滑样条. 这里同样可以使用插值样条的方程, 只需要将所有的 y_j 用 a_j 代替, 因为平滑样条的横坐标 $\{a_j; j=0,\cdots,n\}$ 不需要穿过所有的数据点 $\{y_j; j=0,\cdots,n\}$. 由 (15.16) 式, 有

$$\boldsymbol{H}\boldsymbol{m}=\boldsymbol{u},$$

这里 $\boldsymbol{m}=(m_1,m_2,\cdots,m_{n-1})^{\mathrm{T}}$, 而 $\boldsymbol{u}=(u_1,u_2,\cdots,u_{n-1})^{\mathrm{T}}$. 因为在简单样条情形, $m_0=m_n=0$. 由 (15.15) 式, 向量 $\boldsymbol{u}$ 可以表示为

$$\boldsymbol{u}=\boldsymbol{R}\boldsymbol{a},$$

其中 $\boldsymbol{R}$ 是一个 $(n-1)\times(n+1)$ 的矩阵:

$$\boldsymbol{R}=\begin{bmatrix} r_0 & -(r_0+r_1) & r_1 & 0 & \cdots & \cdots & 0 \\ 0 & r_1 & -(r_1+r_2) & r_2 & 0 & \cdots & 0 \\ \ddots & & \ddots & & \ddots & & \ddots \\ 0 & \cdots & \cdots & 0 & r_{n-2} & -(r_{n-2}+r_{n-1}) & r_{n-1} \end{bmatrix}$$

且

$$\boldsymbol{a}=(a_0,a_1,\cdots,a_n)^{\mathrm{T}},\quad r_j=6h_j^{-1}.$$

所以得到

$$\boldsymbol{H}\boldsymbol{m}=\boldsymbol{R}\boldsymbol{a}. \tag{15.27}$$

这样, 就可以将准则 L 表示为

$$L=p(\boldsymbol{y}-\boldsymbol{a})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{a})+\frac{1}{6}(1-p)\boldsymbol{m}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{m},$$

其中 $\boldsymbol{\Sigma}=\mathrm{diag}\{\sigma_0^2,\sigma_1^2,\cdots,\sigma_n^2\}$, 因为 $\boldsymbol{m}=\boldsymbol{H}^{-1}\boldsymbol{R}\boldsymbol{a}$, 可以将判别准则表示为

$$L=p(\boldsymbol{y}-\boldsymbol{a})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{a})+\frac{1}{6}(1-p)\boldsymbol{a}^{\mathrm{T}}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{H}^{-1}\boldsymbol{R}\boldsymbol{a}.$$

然后可以将每个 $a_0,a_1,\cdots,a_n$ 对 L 求导得到坐标的极值. 以矩阵形式表示, 结果为 (对求导结果再除 2)

$$-p(\boldsymbol{y}-\boldsymbol{a})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}+\frac{1}{6}(1-p)\boldsymbol{a}^{\mathrm{T}}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{H}^{-1}\boldsymbol{R}=\boldsymbol{0},$$

其中 $\mathbf{0}$ 是 $(n+1)\times 1$ 的零向量 $(0,\cdots,0)^{\mathrm{T}}$. 变形后, 可得

$$6p\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{a})=(1-p)\boldsymbol{R}^{\mathrm{T}}\boldsymbol{H}^{-1}\boldsymbol{R}\boldsymbol{a}$$

或

$$6p\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{a})=(1-p)\boldsymbol{R}^{\mathrm{T}}\boldsymbol{m}. \tag{15.28}$$

在等式两边同乘 $\boldsymbol{R\Sigma}$, 得到

$$6p\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{a})=(1-p)\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{m}$$

或

$$6p(\boldsymbol{R}\boldsymbol{y}-\boldsymbol{R}\boldsymbol{a})=(1-p)\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{m}. \tag{15.29}$$

由 $\boldsymbol{H}\boldsymbol{m}=\boldsymbol{R}\boldsymbol{a}$, 可将上式化简为

$$p\boldsymbol{R}\boldsymbol{y}-p\boldsymbol{H}\boldsymbol{m}=\frac{1}{6}(1-p)\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{m}$$

或

$$\left(p\boldsymbol{H}+\frac{1}{6}(1-p)\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R}^{\mathrm{T}}\right)\boldsymbol{m}=p\boldsymbol{R}\boldsymbol{y}. \tag{15.30}$$

这表示由 $n-1$ 个未知数构成的 $n-1$ 个方程, 可以解出 $m_1,m_2,\cdots,m_{n-1}$. 由矩阵形式, (15.30) 式的解可以表示为

$$\boldsymbol{m}=\left(\boldsymbol{H}+\frac{1-p}{6p}\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}\boldsymbol{y}. \tag{15.31}$$

然后通过改写 (15.28) 式, 可以得到 $a_0,a_1,\cdots,a_n$ 的表达式

$$\boldsymbol{a}=\boldsymbol{y}-\frac{1-p}{6p}\boldsymbol{\Sigma}\boldsymbol{R}^{\mathrm{T}}\boldsymbol{m}. \tag{15.32}$$

最后, 将 (15.31) 式代入 (15.32) 式, 可得

$$\boldsymbol{a}=\boldsymbol{y}-\frac{1-p}{6p}\boldsymbol{\Sigma}\boldsymbol{R}^{\mathrm{T}}\left(\boldsymbol{H}+\frac{1-p}{6p}\boldsymbol{R}\boldsymbol{\Sigma}\boldsymbol{R}^{\mathrm{T}}\right)^{-1}\boldsymbol{R}\boldsymbol{y}. \tag{15.33}$$

这样就得到了这个平滑样条的 n 个三次样条的截距值, 其他的系数值可以采用与简单插值样条相同的做法计算出来. 做法是, 像在 15.3 节中讨论的那样采用节点 $\{(x_j,a_j),j=0,\cdots,n\}$ 并同时令 $m_0=m_n=0$. 需要注意的是, 比起简单插值样条, 简单平滑样条唯一增加的计算就是 (15.33) 式.

关于拟合度 F 与平滑度 S 的综合判别值的取值大小可能会有很大的差异. 因此我们不应该对 p 的取值人为地赋予意义 (除非它为 0 或者 1), 较小的 p 值意味

着更大的平滑度, 反之亦然. 在一些应用中可能需要使 p 值非常小, 例如 0.001, 以得到视觉上的明显平滑性. 这也是部分地由于方差的作用, 方差出现在拟合判则的分母中. 较小的方差可以导致拟合度部分比平滑度部分要大得多. 因此, 有必要设定一个非常小的 p 值来强调平滑性.

例 15.9 对表 15-1 的数据, 构建一个简单三次平滑样条. 图 15-8 给出了对死亡率数据构建的简单三次插值样条.

解 图 15-9 与图 15-10 分别给出了 $p=0.5$ 与 $p=0.1$ 的简单三次平滑样条曲线. $p=0.1$ 时平滑样条的系数在表 15-3 中给出. 注意到我们得到的样条与图 15-8 非常相似, 除了在靠近上端的地方, 实际死亡数与观测值有较低的平滑度. 与图 15-10 比较, 可以发现由于对拟合赋予了更小的权重, 图 15-10 的样条平滑度有明显的增加. 标准差的计算与例 15.8 一致, 将所有结果乘以 1000 使得数据更加的合理.[①] □

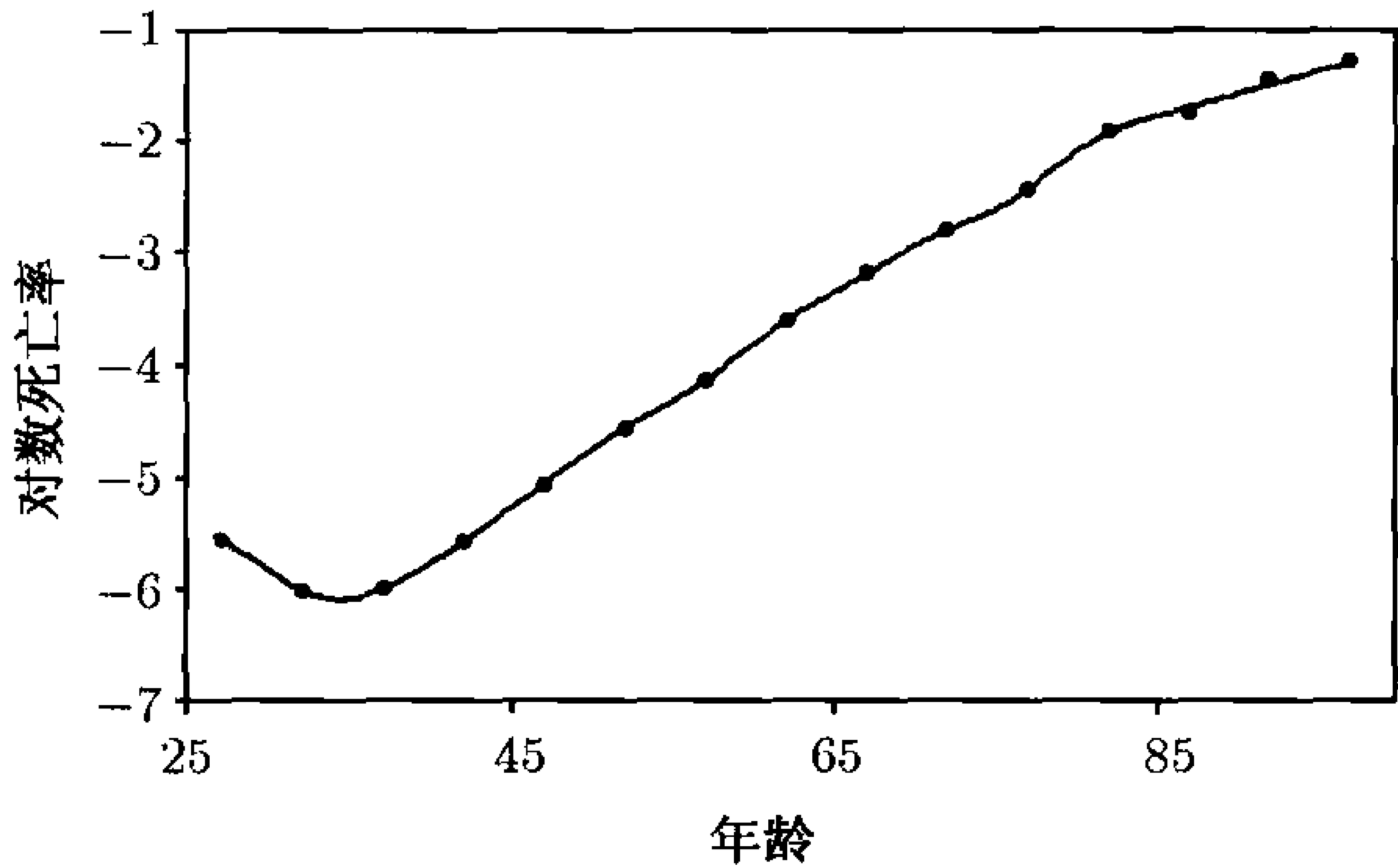


图 15-9 例 15.9 中 $p=0.5$ 下的平滑样条

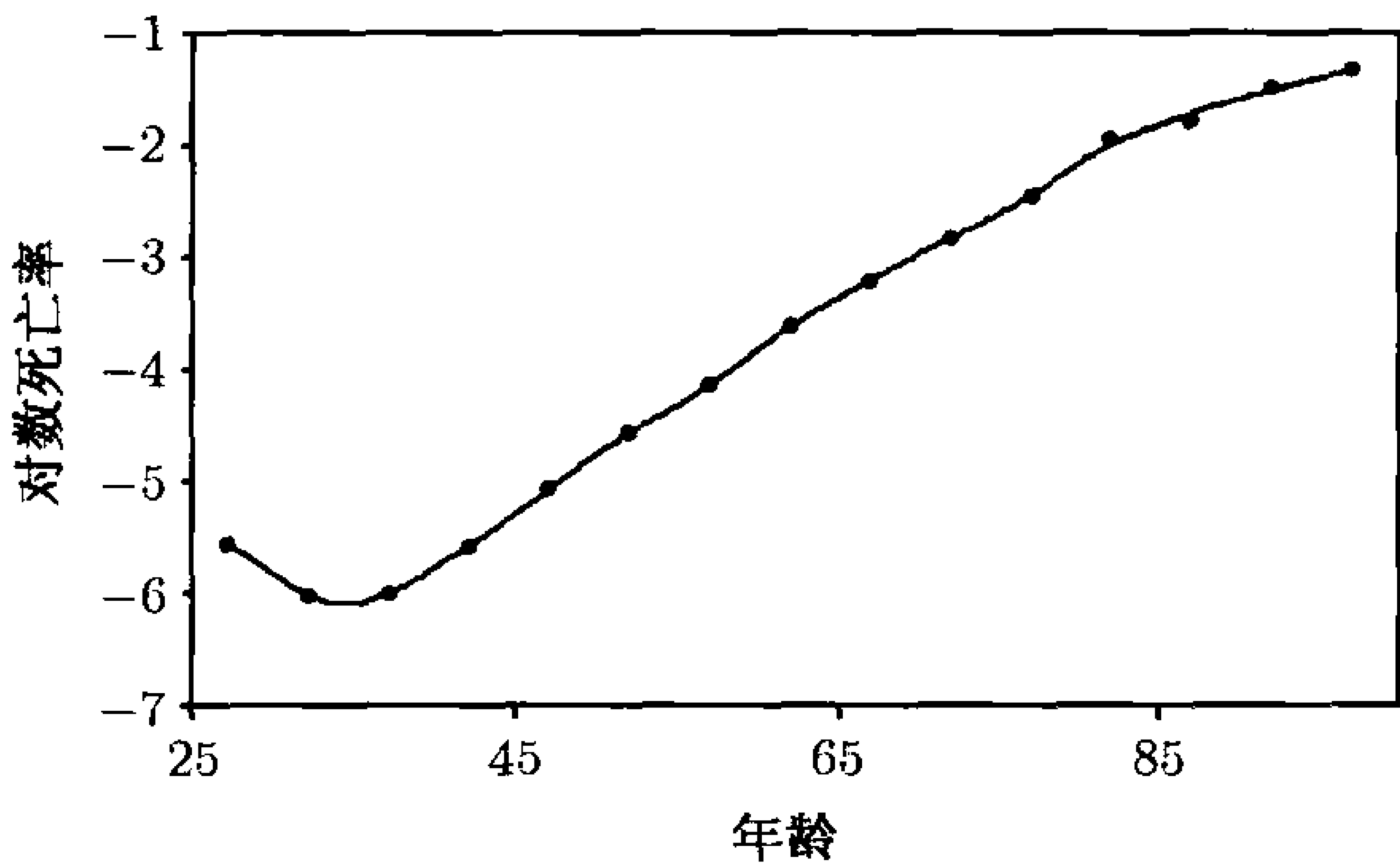


图 15-10 例 15.9 中 $p=0.1$ 下的平滑样条

① 如果不是乘 1 000 的话, 也可以通过改变 p 的值来得到相同的答案. 这里计算标准差的方法不考虑保单规模的影响. 详情可参见 Klugman[75] 的更详细方法. 此处使用的方法暗含了将所有的保单看作是相同规模的, 规模并不重要, 因为有了 1 000 倍的因子, 仍然能被等比例的吸收到 p 中.

表 15-3 例 15.9 中 $p=0.1$ 下的样条系数

j	x_j	a_j	b_j	c_j	d_j
0	27	$3.879\,0 \times 10^{-3}$	$-3.467\,0 \times 10^{-4}$	0	$2.484\,6 \times 10^{-6}$
1	32	$2.456\,0 \times 10^{-3}$	$-1.603\,6 \times 10^{-4}$	$3.726\,9 \times 10^{-5}$	$-8.025\,7 \times 10^{-7}$
2	37	$2.485\,6 \times 10^{-3}$	$1.521\,4 \times 10^{-4}$	$2.523\,0 \times 10^{-5}$	$-5.001\,9 \times 10^{-7}$
3	42	$3.814\,6 \times 10^{-3}$	$3.669\,3 \times 10^{-4}$	$1.772\,8 \times 10^{-5}$	$1.994\,5 \times 10^{-6}$
4	47	$6.341\,7 \times 10^{-3}$	$6.937\,9 \times 10^{-4}$	$4.764\,4 \times 10^{-5}$	$-4.089\,3 \times 10^{-6}$
5	52	$1.049\,1 \times 10^{-2}$	$8.635\,3 \times 10^{-4}$	$-1.369\,5 \times 10^{-5}$	$1.183\,3 \times 10^{-5}$
6	57	$1.594\,5 \times 10^{-2}$	$1.614\,1 \times 10^{-3}$	$1.638\,0 \times 10^{-4}$	$-9.702\,4 \times 10^{-6}$
7	62	$2.689\,8 \times 10^{-2}$	$2.524\,4 \times 10^{-3}$	$1.826\,8 \times 10^{-5}$	$6.263\,3 \times 10^{-6}$
8	67	$4.075\,9 \times 10^{-2}$	$3.176\,9 \times 10^{-3}$	$1.122\,2 \times 10^{-4}$	$-9.443\,5 \times 10^{-7}$
9	72	$5.933\,1 \times 10^{-2}$	$4.228\,2 \times 10^{-3}$	$9.805\,2 \times 10^{-5}$	$3.973\,7 \times 10^{-5}$
10	77	$8.789\,1 \times 10^{-2}$	$8.189\,0 \times 10^{-3}$	$6.941\,1 \times 10^{-4}$	$-6.680\,4 \times 10^{-5}$
11	82	$1.378\,4 \times 10^{-1}$	$1.012\,0 \times 10^{-2}$	$-3.079\,4 \times 10^{-4}$	$2.157\,2 \times 10^{-5}$
12	87	$1.834\,4 \times 10^{-1}$	$8.658\,3 \times 10^{-3}$	$1.563\,3 \times 10^{-5}$	$-1.028\,2 \times 10^{-6}$
13	92	$2.269\,9 \times 10^{-1}$	$8.737\,6 \times 10^{-3}$	$2.102\,1 \times 10^{-7}$	$-1.401\,4 \times 10^{-8}$

例 15.9 说明了如何用平滑样条来自动地同时进行插值与平滑处理. 每 5 个年龄的节点通过 (15.32) 式来平滑. 调整后的节点再用来作为一个插值样条的节点. 此处插值用到的数据其实是各区间中点年龄调整后的死亡率. 例 15.9 的平滑效果在视觉上不是很明显, 因为原始数据本身已经相当平滑了. 下个例子说明了对一列非常强噪声的数据我们如何通过平滑样条来进行显著地平滑处理.

例 15.10 表 15-4 给出了 15 年的观测死亡率. 数据来自 Miller[94] 第 11 页, 由图 15-11 表示. 通过改变 p 值拟合平滑样条直至出现合理的平滑度, 并给出各年龄的修订死亡率.

解 与例 15.9 不同的是, 这里的个数代表死亡人数, 而不再是金额, 可以直接估计死亡率的方差 (见例 15.8). 为了方便, 将标准差乘因子 10. 出于保险的考虑, 我们对样条在节点处的取值 a_j 更感兴趣. 表 15-4 给出了插值且 $p=0.5, 0.1, 0.05$ 下的样条值由图 15-12 至图 15-14 表示. 注意到对于 $p=0.5$, 平滑性相当明显, 但是一些个别点对于结果仍然有着很大的影响. 例如 76 岁出现的大量实际死亡人数就使得曲线被拉高. 通过减少 p 值, 可以得到更大的平滑度, 这一点可以由图 15-12 至图 15-14 明显看出. □

例 15.10 阐述了样条的平滑能力. 然而, 我们仍然面临 p 值的选择. 在实际中, 这是由专业判断与视觉观测决定的. 如例 15.9, 数据点集很大且观测数据已经具有一定程度的平滑性, 那么我们更加需要一个更靠近观测数据的拟合曲线. 如例 15.10

得到的数据比较有限, 那么就需要一定程度的平滑, 这时主观判断很重要. 对于任何规模的数据, 都可以进行正式的拟合优度检验. 拟合标准 F 服从 $n+1$ 个自由度的卡方分布, 这可以起到一些指示作用. 其他的拟合检验例如游程 (runs) 检验可以用来判别拟合样条的一些特殊异常之处.

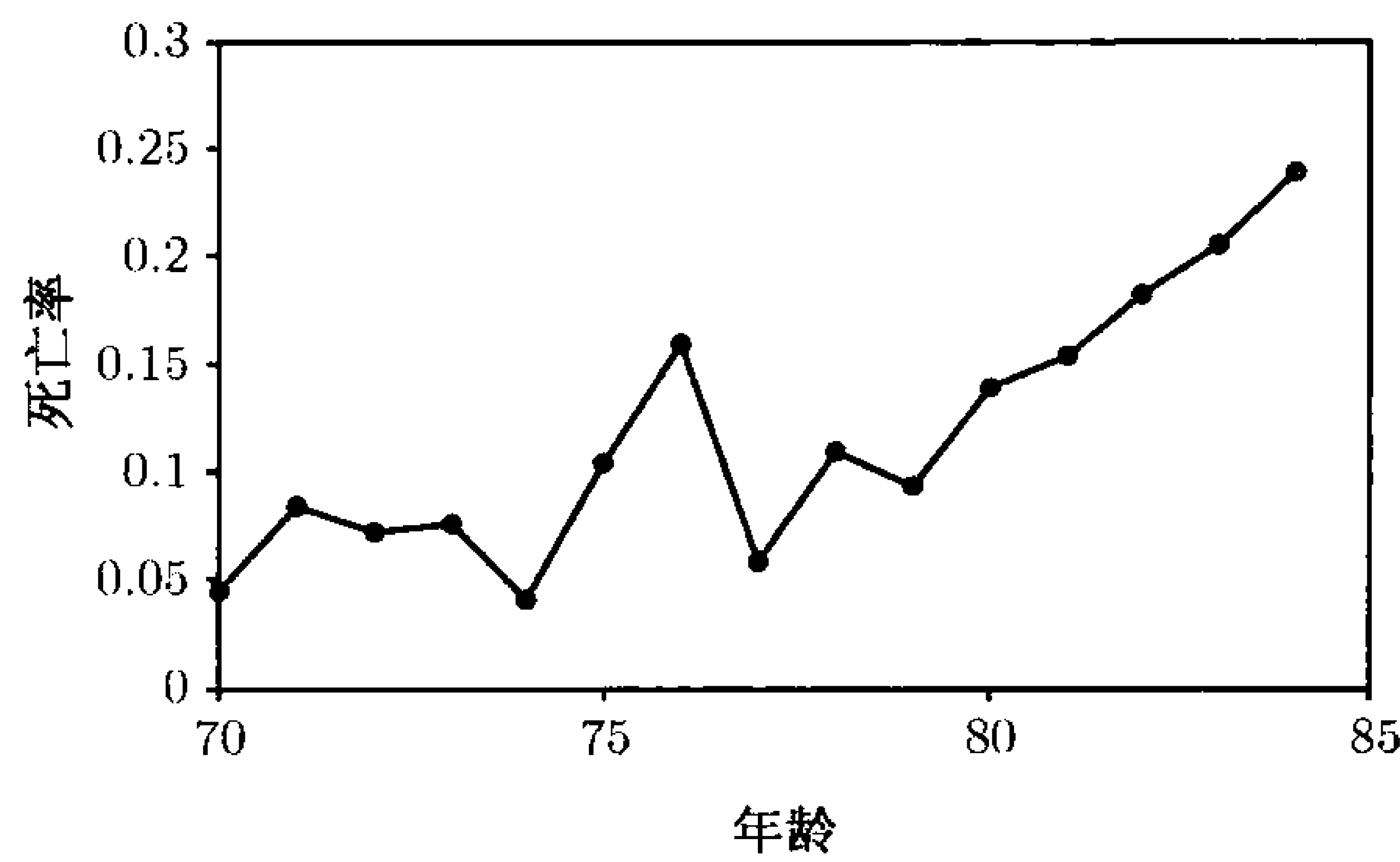


图 15-11 例 15.10 的死亡率数据

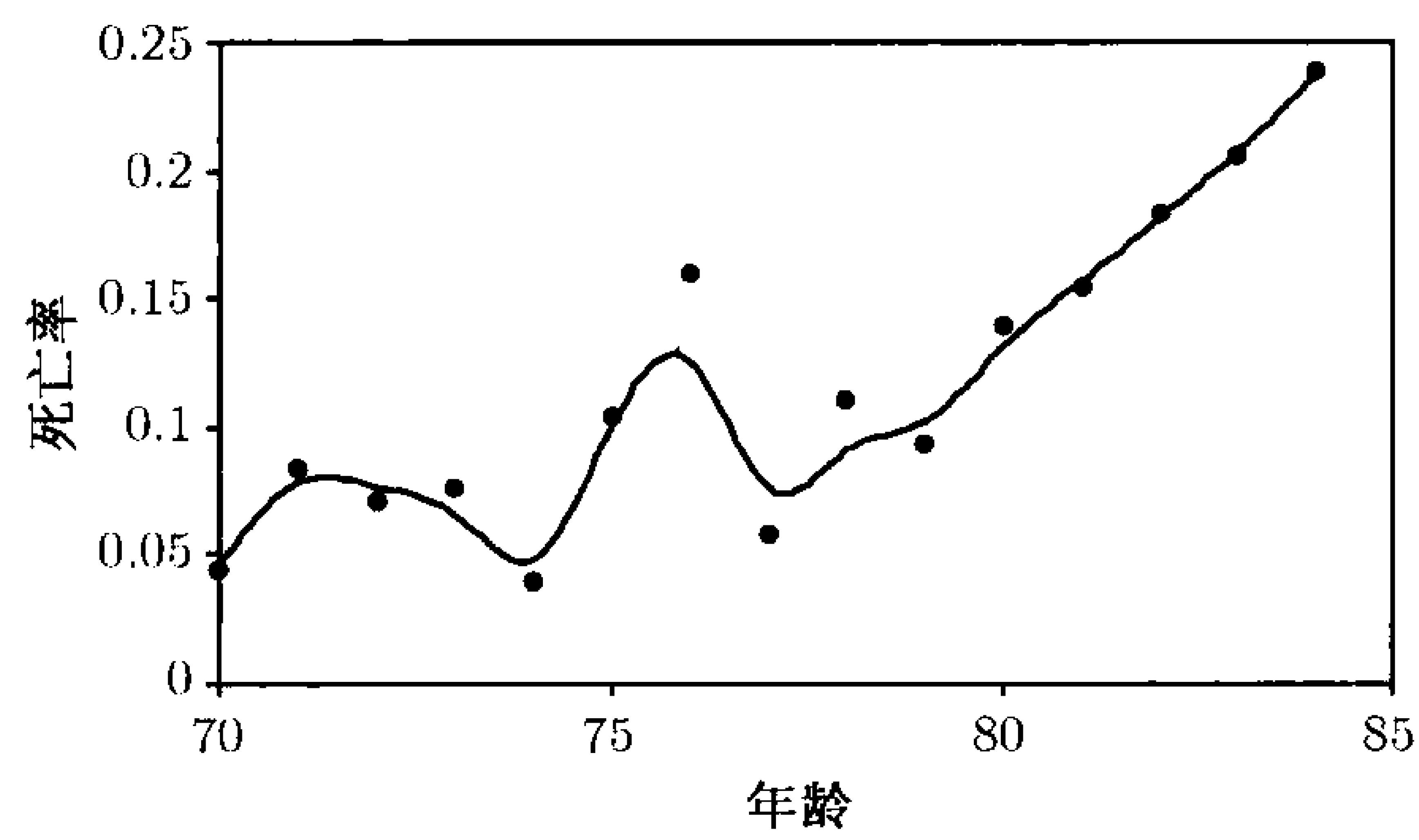


图 15-12 $p = 0.5$ 下的死亡率数据的平滑样条

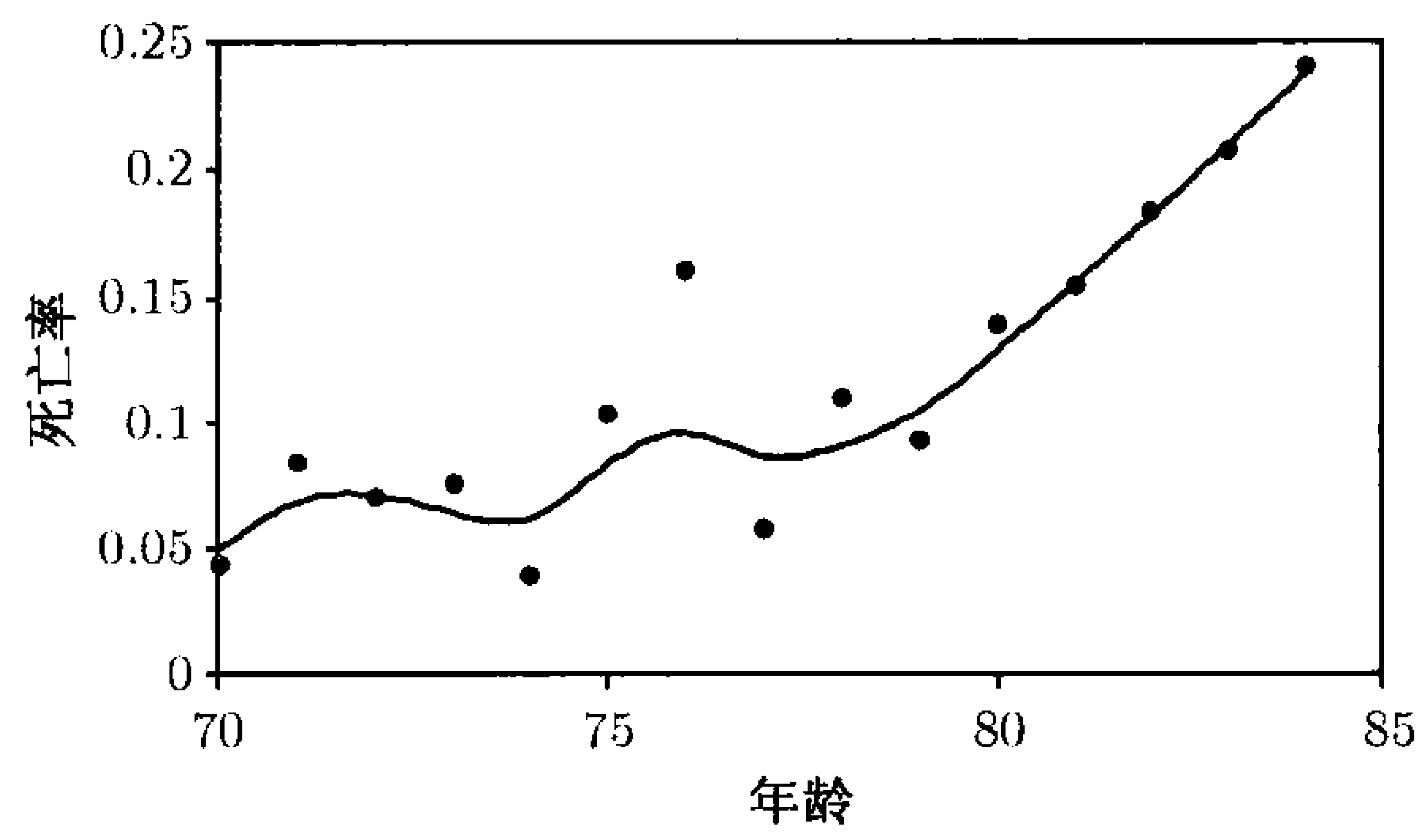


图 15-13 $p = 0.1$ 下的死亡率数据的平滑样条

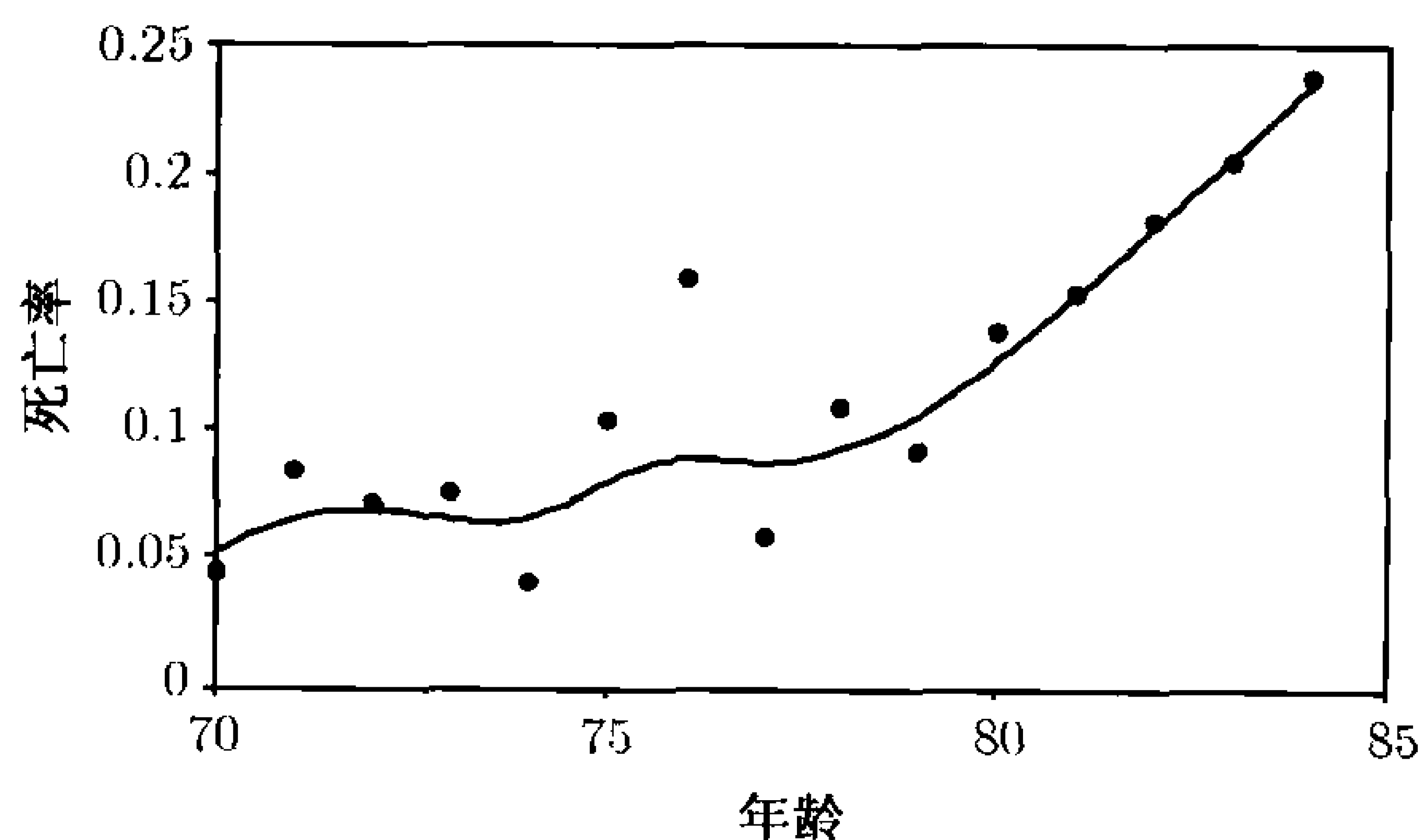


图 15-14 $p = 0.05$ 下的死亡率数据的平滑样条

表 15-4 例 15.10 中的死亡率以及插值

j	年龄 x_j	风险暴露	死亡人数	死亡率估计	光滑化因子		
					$p = 0.5$	$p = 0.1$	$p = 0.05$
0	70	135	6	0.044	0.046	0.050	0.052
1	71	143	12	0.084	0.078	0.069	0.065
2	72	140	10	0.071	0.077	0.071	0.069
3	73	144	11	0.076	0.066	0.064	0.065
4	74	149	6	0.040	0.049	0.062	0.066
5	75	154	16	0.104	0.100	0.084	0.080
6	76	150	24	0.160	0.126	0.096	0.089
7	77	139	8	0.058	0.076	0.087	0.088
8	78	145	16	0.110	0.091	0.091	0.093
9	79	140	13	0.093	0.102	0.105	0.107
10	80	137	19	0.139	0.131	0.128	0.128
11	81	136	21	0.154	0.157	0.155	0.154
12	82	126	23	0.183	0.182	0.181	0.181
13	83	126	26	0.206	0.208	0.209	0.208
14	84	109	26	0.239	0.238	0.237	0.236
总和		2 073	237				

习题

15.11 考虑对点 (0,0),(1,2),(2,1),(3,3) 进行简单三次平滑样条处理, $p = 0.9$, 标准差为 0.5.

(a) 用 (15.33) 式求解节点处的截距值.

(b) 用 (15.16) 式与 (15.18) 式求解简单三次平滑样条.

(c) 给出从 $x = -0.5$ 到 $x = 2.5$ 的样条图像.

第16章 信度理论

16.1 引 言

信度理论是一种定量方法, 保险人可借助这种方法对单个或一组风险进行全面的经验费率厘定 (即根据承保经验来调整未来的保费). 如果某个投保人的表现始终优于费率手册的假设 (有时称之为**净保费**), 则该投保人可能会要求降低保费.

投保人提出这种要求的理由如下: 手册费率反映整个费率级别的预期表现, 并且假定每个等级的所有风险是同质的. 但是, 不存在完美的费率系统, 即使考虑了所有的核保标准, 也会存在异质的风险. 因此, 某些投保人将具有低于费率手册的风险. 同理, 对那些有更高风险的投保人应当加收保费, 然而投保人绝不会主动提出这样的要求! 尽管如此, 从公平和经济的角度看, 增加保费或许是必要的.

因此保险人不得不面对如下的问题: 投保人的经验数据与预期的差异在多大程度上可以归结为索赔额的随机波动, 又有多大程度是因为投保人自身的风险水平确实高于或低于平均水平? 也就是说, 投保人的经验数据有多大的可信度? 这里有两个基本的考虑.

(1) 保险人掌握某个投保人的信息越多, 该投保人经验数据的*可信度*就应当越大, 反之亦然. 同理, 在团体保险中, 较大群体的可信度比较小群体的可信度更大.

(2) 出于竞争的原因, 保险人不得不给予某类特定投保人完全可信度 (即收取的保费完全基于该投保人的经验数据而不是费率手册) 或是近乎完全的可信度, 以便维系该类业务.

信度理论的另一个应用是设计分级费率 (classification system). 例如, 在工伤保险中可能包含许多的职业, 其中有些职业的数据可能非常有限. 为了精确地估计每个类别的保险成本, 可以将有限的经验数据和其他信息结合起来, 比如过去的费率, 或与其紧密相关的职业的数据, 等等.

从统计的观点看, 信度理论可能会得到一些违反直觉的结果. 如果得到了某被保险人或被保险人团体的历史数据, 受过的统计训练会促使我们使用样本均值或是其他无偏估计量. 而信度理论却告诉我们, 更好的方法是只给予这类估计量部分的权重, 而剩下的权重将赋予其他信息的估计量. 可以发现无偏性的牺牲换来了均方误差的减少.

信度理论让保险人能够定量地分析上述问题, 本章将主要介绍这种理论. 16.2

节将回顾相关的统计概念, 在 9.2 节和 12.4 节中提到的一些概念也会再次重复, 同时也将介绍一些新的概念和公式.

16.3 节介绍**有限波动信度理论**. 它是在二十世纪早期发展起来的方法, 提供了一种机制确定投保人的经验数据是完全信度 (16.3.1 节) 还是部分 (16.3.2 节) 信度. 尽管该方法缺少充分的数学理论支持, 但这个理论提供了一种创造性的处理方法, 现在仍然有用武之地.

Bühlmann 的经典文章 [18] 中提出了一种统计学的背景框架, 为信度理论的建立和发展提供了理论基础. 这种称为**最大精度信度理论**①是由 Bühlmann 正式提出的, 虽然在此之前已经出现了一段时间. 16.4 节会介绍这种方法. 16.4.4 节将讨论最简单的模型 (Bühlmann[18]). 该模型的改进由 Bühlmann 和 Straub([20]) 提出, 该模型将在 16.4.5 节讨论. 16.4.6 节将提出严格信度的概念.

信度理论在进行实际应用时需要根据数据对未知模型进行估计参数. 非参数估计 (没有特定模型的限制, 并且都是一般性的参数, 如均值和方差等) 将在 16.5.1 节讨论, 半参数估计 (有些参数由特定的先验分布决定) 在 16.5.2 节讨论, 最后将在 16.5.3 节讨论完全参数估计 (所有参数都来自某些特定分布).

我们以 Arthur Bailey1950[8] 中第 8 页的一段话来结束引言, 它恰当地描述了信度理论的发展过程. 在这里我们也要向早期的精算师表示敬意, 因为在只有一些简单的数学工具的条件下他们却也能够提出一些行之有效的公式, 并且与在本章通过细致推导所获得的公式相差无几.

> 讨论到此, 一般人不得不承认, 精算师们的观点背后似乎存在着一些模糊的逻辑推理, 非常地含糊不清, 他们很难理解这些观点. 一个训练有素的统计学家会大喊: “荒唐! 竟然直接违背公认的统计估计理论”. 精算师们也承认他们使用了一些在数学上没有被严格证明的结论, 当中所包含的价值标准都建立在主观判断的基础之上, 唯一能够证实的是在实际应用中, 这些方法确实有效. 千万不要忘记, 它们已经被验证了许多次, 而且确实有效!

16.2 统计学概念

本节将提出各种与信度理论有关的统计概念. 因为大部分内容都是回顾性的, 所以若读者拥有好的统计背景, 这些内容可以略过. 当然, 也有一些材料或许是读者并不了解的, 因此也不必完全跳过这一节. 后续章节也将涉及本节的相关内容.

16.2.1 条件分布

假定 X 和 Y 是两个随机变量, 联合概率函数 (pf) 或概率密度函数为

① 有限波动和最大精度的术语至少可以追溯到 1943 年 Arthur Bailey[7].

(pdf)[1] $f_{X,Y}(x,y)$, 边缘概率函数为 $f_X(x)$ 和 $f_Y(y)$. 给定 $Y=y$ 时 X 的条件概率函数为

$$f_{X|Y}(x|y)=\frac{f_{X,Y}(x,y)}{f_Y(y)}. \tag{16.1}$$

如果 X 和 Y 是离散型随机变量, 则 (16.1) 式就是在假定 $Y=y$ 的条件下 $X=x$ 的条件概率. 如果 X 和 Y 是连续型随机变量, 则 (16.1) 式就是条件概率密度的定义. 当 X 和 Y 为独立随机变量时, 有

$$f_{X,Y}(x,y)=f_X(x)f_Y(y),$$

因此 (16.1) 变为

$$f_{X|Y}(x|y)=f_X(x),$$

这时, X 的条件分布和边缘分布相同.

例 16.1 假定 X 和 Z 是独立 Poisson 随机变量, 均值分别为 λ_1 和 λ_2. 设 $Y=X+Z$, 证明 $X|Y=y$ 为二项分布, 参数为 $m=y$, $q=\lambda_1/(\lambda_1+\lambda_2)$(见 [58], p131).

证明 给定 $Y=y$, X 的条件分布是

$$\begin{aligned}
f_{X|Y}(x|y)&=\frac{f_{X,Y}(x,y)}{f_Y(y)}=\frac{\Pr(X=x,Y=y)}{\Pr(Y=y)}\\
&=\frac{\Pr(X=x,Z=y-x)}{\Pr(Y=y)}=\frac{\Pr(X=x)\Pr(Z=y-x)}{\Pr(Y=y)}\\
&=\frac{\dfrac{\lambda_1^x\mathrm{e}^{-\lambda_1}}{x!}\dfrac{\lambda_2^{y-x}\mathrm{e}^{-\lambda_2}}{(y-x)!}}{\dfrac{(\lambda_1+\lambda_2)^y\mathrm{e}^{-\lambda_1-\lambda_2}}{y!}}=\frac{y!}{x!(y-x)!}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^x\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{y-x},
\end{aligned}$$

对 $x=0,1,2,\cdots,y$ 均成立. 这是一个参数为 $m=y$, $q=\lambda_1/(\lambda_1+\lambda_2)$ 的二项分布. □

注意到 (16.1) 式可以改写为

$$f_{X,Y}(x,y)=f_{X|Y}(x|y)f_Y(y), \tag{16.2}$$

说明联合分布可以由边缘分布和条件分布的乘积得到. 由于 X 的边缘分布可以通过对联合分布中的 y 求积分 (或求和) 得到

$$f_X(x)=\int f_{X,Y}(x,y)\mathrm{d}y,$$

---

① 在没有明确说明时随机变量可能是连续型、离散型或混合型, 将使用术语概率函数及缩写 pf.. 只有当随机变量是连续型时, 使用术语概率密度函数及其缩写 pdf.

再利用 (16.2) 式, 有

$$f_X(x) = \int f_{X|Y}(x|y) f_Y(y) \mathrm{d}y. \tag{16.3}$$

有趣的是, 公式 (16.3) 可以看成是一个混合分布 (见 4.4.5 节). 为了说明这一点. 假设条件分布 $f_{X|Y}(x|y)$ 是一个常见的参数分布, 而 y 是分布函数为 $f_Y(y)$ 的随机变量 Y 的一次实现. 在 4.6.3 节中证明了, 给定 $\Theta=\theta$ 的条件下, X 服从均值为 θ 的 Poisson 分布, 同时 Θ 服从 gamma 分布, 则 X 的边缘分布是负二项分布. 并且, 例 4.30 证明了, 若 $X|\Theta$ 服从均值为 Θ、方差为 v 的正态分布, 而 Θ 服从均值为 μ、方差为 a 的正态分布, 则 X 的边缘分布是正态分布, 且均值为 μ、方差为 $a+v$.

另外, 由于 (16.2) 式中 X 和 Y 地位的对称性, 易见

$$f_{X|Y}(x|y) f_Y(y) = f_{Y|X}(y|x) f_X(x),$$

这是因为上式两边都等于 X 和 Y 的联合分布. 两边再同时用 $f_Y(y)$ 去除就得到贝叶斯定理

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)}.$$

16.2.2　条件期望

如前所述, 假定 X 和 Y 是两个随机变量, 且给定 $Y=y$ 的条件下, X 的条件 pf 是 $f_{X|Y}(x|y)$. 显然, 这的确是一个概率分布, 它的均值可以表示为

$$\mathrm{E}(X|Y=y) = \int x f_{X|Y}(x|y) \mathrm{d}x, \tag{16.4}$$

若是离散分布, 则积分号用求和号代替. 易见 (16.4) 式是 y 的函数, 通常在 (16.4) 式的右边将 y 用 Y 代换, 然后将这个条件期望看作和 Y 有关的随机变量. 因此, 可以将 (16.4) 式的左边替换为 $\mathrm{E}(X|Y)$, 则 $\mathrm{E}(X|Y)$ 自身也是一个随机变量, 它是随机变量 Y 的一个函数. $\mathrm{E}(X|Y)$ 的期望为

$$\mathrm{E}[\mathrm{E}(X|Y)] = \mathrm{E}(X). \tag{16.5}$$

为证明这一点, 由 (16.3) 式和 (16.4) 式可得

$$\begin{aligned}\mathrm{E}[\mathrm{E}(X|Y)] &= \int \mathrm{E}(X|Y=y) f_Y(y) \mathrm{d}y = \iint x f_{X|Y}(x|y) \mathrm{d}x f_Y(y) \mathrm{d}y \\ &= \int x \int f_{X|Y}(x|y) f_Y(y) \mathrm{d}y \mathrm{d}x = \int x f_X(x) \mathrm{d}x = \mathrm{E}(X).\end{aligned}$$

对离散随机变量情形, 证明类似.

例 16.2　利用条件期望推导负二项分布的均值. 即若 $X|\Theta \sim Poisson(\Theta)$, 且 $\Theta \sim gamma(\alpha,\beta)$, 则 $X \sim$ 负二项分布 $r=\alpha, \beta=\beta$.

解 因为有

$$E(X|\Theta) = \Theta,$$

所以

$$E(X) = E[E(X|\Theta)] = E(\Theta).$$

从附录 A 可知 gamma 分布 Θ 的均值为 $\alpha\beta$, 故 $E(X) = \alpha\beta$. □

为方便起见, 可以将 (16.4) 式中的 X 换为任意函数 $h(X, Y)$, 得到更一般的定义

$$E[h(X,Y)|Y=y] = \int h(x,y) f_{X|Y}(x|y)\mathrm{d}x.$$

类似地, $E[h(X,Y)|Y]$ 也是一个条件期望, 同时也可以看作是随机变量 Y 的函数. 于是 (16.5) 式可以更一般地表示为

$$E\{E[h(X,Y)|Y]\} = E[h(X,Y)]. \tag{16.6}$$

为证明这一点, 从 (16.2) 式得

$$\begin{aligned}
E\{E[h(X,Y)|Y]\} &= \int E[h(X,Y)|Y=y] f_Y(y)\mathrm{d}y \\
&= \iint h(x,y) f_{X|Y}(x|y)\mathrm{d}x f_Y(y)\mathrm{d}y \\
&= \iint h(x,y)[f_{X|Y}(x|y) f_Y(y)]\mathrm{d}x\mathrm{d}y \\
&= \iint h(x,y) f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y \\
&= E[h(X,Y)].
\end{aligned}$$

若 $h(X,Y) = [X - E(X|Y)]^2$, 则它的期望值是给定 Y 的条件下 X 的条件分布的方差

$$\mathrm{Var}(X|Y) = E\{[X - E(X|Y)]^2|Y\}. \tag{16.7}$$

显然, (16.7) 式仍然是随机变量 Y 的函数.

更有启发性的是利用两个随机变量 X 和 Y 计算 X 的方差. 注意 (16.7) 式可以表示为

$$\mathrm{Var}(X|Y) = E(X^2|Y) - [E(X|Y)]^2.$$

因此有

$$\begin{aligned}
E[\mathrm{Var}(X|Y)] &= E\{E(X^2|Y) - [E(X|Y)]^2\} \\
&= E[E(X^2|Y) - E\{[E(X|Y)]^2\} \\
&= E(X^2) - E\{[E(X|Y)]^2\}.
\end{aligned}$$

同时由于 $\mathrm{Var}[h(Y)] = \mathrm{E}\{[h(Y)]^2\} - \{\mathrm{E}[h(Y)]\}^2$, 令 $h(Y) = \mathrm{E}(X|Y)$ 可得

$$\begin{aligned}\mathrm{Var}[\mathrm{E}(X|Y)] &= \mathrm{E}\{[\mathrm{E}(X|Y)]^2\} - \{\mathrm{E}[\mathrm{E}(X|Y)]\}^2 \\ &= \mathrm{E}\{[\mathrm{E}(X|Y)]^2\} - [\mathrm{E}(X)]^2,\end{aligned}$$

进而有

$$\begin{aligned}\mathrm{E}[\mathrm{Var}(X|Y)] + \mathrm{Var}[\mathrm{E}(X|Y)] &= \mathrm{E}(X^2) - \mathrm{E}\{[\mathrm{E}(X|Y)]^2\} + \mathrm{E}\{[\mathrm{E}(X|Y)]^2\} - [\mathrm{E}(X)]^2 \\ &= \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = \mathrm{Var}(X).\end{aligned}$$

从而得到重要公式

$$\mathrm{Var}(X) = \mathrm{E}[\mathrm{Var}(X|Y)] + \mathrm{Var}[\mathrm{E}(X|Y)]. \tag{16.8}$$

公式 (16.8) 式表明, X 的方差由两部分组成: 条件方差的期望与条件期望的方差之和.

例 16.3 *推导负二项分布的方差.*

解 Poisson 分布具有相等的均值和方差, 也就是说

$$\mathrm{E}(X|\Theta) = \mathrm{Var}(X|\Theta) = \Theta,$$

并且, 由 (16.8) 式得

$$\mathrm{Var}(X) = \mathrm{E}[\mathrm{Var}(X|\Theta)] + \mathrm{Var}[\mathrm{E}(X|\Theta)] = \mathrm{E}(\Theta) + \mathrm{Var}(\Theta).$$

因为 Θ 服从 $\mathrm{gamma}(\alpha,\beta)$ 分布, 且 $\mathrm{E}(\Theta) = \alpha\beta$, $\mathrm{Var}(\Theta) = \alpha\beta^2$. 故负二项分布的方差为

$$\mathrm{Var}(X) = \mathrm{E}(\Theta) + \mathrm{Var}(\Theta) = \alpha\beta + \alpha\beta^2 = \alpha\beta(1+\beta).$$ □

例 16.4 *例 4.30 证明了, 若 $X|\Theta$ 服从均值为 Θ、方差为 v 的正态分布, 而 Θ 服从均值为 μ、方差为 a 的正态分布, 则 X(无条件地) 为正态分布, 且均值为 μ、方差为 $a+v$. 利用 (16.5) 式和 (16.8) 式直接计算 X 的均值和方差.*

解 先计算均值

$$\mathrm{E}(X) = \mathrm{E}[\mathrm{E}(X|\Theta)] = \mathrm{E}(\Theta) = \mu,$$

再计算方差

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}[\mathrm{Var}(X|\Theta)] + \mathrm{Var}[\mathrm{E}(X|\Theta)] \\ &= \mathrm{E}(v) + \mathrm{Var}(\Theta) = v + a,\end{aligned}$$

由于 v 是常数, 故 $\mathrm{E}(v) = v$. □

例 16.5 *考虑 Poisson 参数为 λ 的复合 Poisson 分布 $X = Y_1 + Y_2 + \cdots + Y_N$, 而且 $\mathrm{E}(Y_i) = \mu_Y$ 和 $\mathrm{Var}(Y_i) = \sigma_Y^2$. 计算 X 的均值和方差.*

解 在第 6 章中用公式 (16.8) 可得到结果

$$\mathrm{E}(X) = \lambda\mu_Y \quad 和 \quad \mathrm{Var}(X) = \lambda(\mu_Y^2 + \sigma_Y^2). \qquad \square$$

16.2.3 非参数型无偏估计量

曾经在 9.2.2 节中论述过无偏估计, 它在信度理论的发展中起着重要的作用. 下面证明两个常用的估计量是无偏的.

定理 16.6 *若 $X_1, \cdots, X_N$ 是一列相互独立但不一定同分布的随机变量, 有共同的均值 $\mu = \mathrm{E}(X_j)$ 和方差 $v = \mathrm{Var}(X_j)$, 则*

$$\bar{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$$

为 μ 的无偏估计量, 且

$$\hat{v} = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \bar{X})^2 \tag{16.9}$$

为 v 的无偏估计量.

证明 对 $\bar{X}$, 有

$$\mathrm{E}(\bar{X}) = \mathrm{E}\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right) = \frac{1}{n}\sum_{j=1}^{n}\mathrm{E}(X_j) = \mu.$$

至于方差的估计, 先给出一个有用的结论:

$$\begin{aligned}
\sum_{j=1}^{n}(X_j - \bar{X})^2 &= \sum_{j=1}^{n}(X_j - \mu + \mu - \bar{X})^2 \\
&= \sum_{j=1}^{n}(X_j - \mu)^2 + 2\sum_{j=1}^{n}(X_j - \mu)(\mu - \bar{X}) + \sum_{j=1}^{n}(\mu - \bar{X})^2 \\
&= \sum_{j=1}^{n}(X_j - \mu)^2 + 2(\mu - \bar{X})\sum_{j=1}^{n}(X_j - \mu) + n(\mu - \bar{X})^2 \\
&= \sum_{j=1}^{n}(X_j - \mu)^2 + 2(\mu - \bar{X})n(\bar{X} - \mu) + n(\mu - \bar{X})^2 \\
&= \sum_{j=1}^{n}(X_j - \mu)^2 - n(\bar{X} - \mu)^2.
\end{aligned} \tag{16.10}$$

利用 $X_1, \cdots, X_n$ 的独立性可得

$$\operatorname{Var}(\bar{X}) = \operatorname{Var}\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right) = \frac{1}{n^2}\sum_{j=1}^{n}\operatorname{Var}(X_j) = \frac{1}{n^2}\sum_{j=1}^{n} v = \frac{v}{n}.$$

再对 (16.10) 式两边取期望

$$\begin{aligned}
\mathrm{E}\left[\sum_{j=1}^{n}(X_j - \bar{X})^2\right] &= \mathrm{E}\left[\sum_{j=1}^{n}(X_j - \mu)^2\right] - n\mathrm{E}[(\bar{X} - \mu)^2] \\
&= \sum_{j=1}^{n}\mathrm{E}[(X_j - \mu)^2] - n\operatorname{Var}(\bar{X}) \\
&= \sum_{j=1}^{n}\operatorname{Var}(X_j) - n\frac{v}{n} = \left(\sum_{j=1}^{n} v\right) - v \\
&= (n-1)v.
\end{aligned}$$

等式两边同时除以 $n-1$ 即可证明 $\hat{v}$ 是 v 的无偏估计. □

下面的例子是上述结论的一般化推广, 虽然看上去有一些人为构造的痕迹, 但是与将在 16.4.5 节中介绍的 Bühlmann-Straub 模型具有重要的联系.

例 16.7　*假设 $X_1, \cdots, X_n$ 是一列相互独立的随机变量, 具有共同的均值 $\mu = \mathrm{E}(X_j)$ 和方差 $\operatorname{Var}(X_j) = \beta + \alpha/m_j$, $\alpha, \beta > 0$, 且所有的 $m_j \geqslant 1$. 令 $m = \sum_{j=1}^{n} m_j$, 考虑 3 个估计量*

$$\bar{X} = \frac{1}{m}\sum_{j=1}^{n} m_j X_j, \quad \hat{\mu}_1 = \frac{1}{m}\sum_{j=1}^{n} X_j,$$

和

$$\hat{\mu}_2 = \frac{\sum_{j=1}^{n}\frac{m_j X_j}{m_j\beta+\alpha}}{\sum_{j=1}^{n}\frac{m_j}{m_j\beta+\alpha}}.$$

证明 3 个估计量均为 μ 的无偏估计, 并将它们按均方误差的大小排序. 同时推导出一个平方和式的期望值, 该期望值有助于估计 α 和 β.

证明　首先考虑 $\bar{X}$.

$$\mathrm{E}(\bar{X}) = m^{-1}\sum_{j=1}^{n} m_j\mathrm{E}(X_j) = m^{-1}\sum_{j=1}^{n} m_j\mu = \mu,$$

$$\operatorname{Var}(\bar{X}) = m^{-2}\sum_{j=1}^{n} m_j^2\operatorname{Var}(X_j) = m^{-2}\sum_{j=1}^{n} m_j^2\left(\beta + \frac{\alpha}{m_j}\right)$$

$$= \alpha m^{-1} + \beta m^{-2} \sum_{j=1}^{n} m_j^2.$$

定理 16.6 中已证明估计量 $\hat{\mu}_1$ 是无偏的. 同时可以得到

$$\mathrm{Var}(\hat{\mu}_1) = n^{-2} \sum_{j=1}^{n} \mathrm{Var}(X_j) = n^{-2} \sum_{j=1}^{n} \left(\beta + \frac{\alpha}{m_j}\right)$$
$$= \beta n^{-1} + n^{-2} \alpha \sum_{j=1}^{n} m_j^{-1}.$$

至于 $\hat{\mu}_2$, 有

$$\mathrm{E}(\hat{\mu}_2) = \frac{\sum_{j=1}^{n} \frac{m_j}{m_j\beta+\alpha} \mathrm{E}(X_j)}{\sum_{j=1}^{n} \frac{m_j}{m_j\beta+\alpha}} = \frac{\sum_{j=1}^{n} \frac{m_j}{m_j\beta+\alpha} \mu}{\sum_{j=1}^{n} \frac{m_j}{m_j\beta+\alpha}} = \mu$$

以及

$$\mathrm{Var}(\hat{\mu}_2) = \frac{\sum_{j=1}^{n} \left(\frac{m_j}{m_j\beta+\alpha}\right)^2 \left(\beta + \frac{\alpha}{m_j}\right)}{\left(\sum_{j=1}^{n} \frac{m_j}{m_j\beta+\alpha}\right)}$$
$$= \frac{\sum_{j=1}^{n} \left(\frac{m_j}{m_j\beta+\alpha}\right)}{\left(\sum_{j=1}^{n} \frac{m_j}{m_j\beta+\alpha}\right)^2} = \left(\sum_{j=1}^{n} \frac{m_j}{m_j\beta+\alpha}\right)^{-1}.$$

现在考虑对估计量的方差进行排序. (因为所有估计量都是无偏的, 所以它们的均方误差就是方差, 要比较均方误差只需比较方差). 为了证明 $\mathrm{Var}(\hat{\mu}_1)$ 与 $\mathrm{Var}(\bar{X})$ 无法确定大小, 计算两者的差

$$\mathrm{Var}(\bar{X}) - \mathrm{Var}(\hat{\mu}_1) = \alpha \left(m^{-1} - n^{-2} \sum_{j=1}^{n} m_j^{-1}\right) + \beta \left(m^{-2} \sum_{j=1}^{n} m_j^2 - n^{-1}\right).$$

注意到

$$\frac{1}{n} \sum_{j=1}^{n} m_j^2 \geqslant \left(\frac{1}{n} \sum_{j=1}^{n} m_j\right)^2 = \frac{m^2}{n^2}$$

(不等号左边是样本的二阶矩, 最右边是样本均值的平方), 然后两边同时乘上 nm^{-2}. 故 β 的系数非负. 下面再说明 α 的系数非正. 注意到

$$\frac{n}{\sum_{j=1}^{n} m_j^{-1}} \leqslant \frac{1}{n} \sum_{j=1}^{n} m_j = \frac{m}{n}$$

(调和平均总是不大于算术平均) 将上式两边同时乘上 n, 然后同时取倒数即可. 综上所述, 只要选取合适的 α 和 β, 这个差值可以是正数, 也可以是负数.

下面证明一个更强的关于 $\hat{\mu}_2$ 的结论. 考虑形如 $\hat{\mu}=\sum_{j=1}^{n} a_j X_j$ 的估计量, 其中 $\sum_{j=1}^{n} a_j = 1$(为了保证无偏性). 前述 3 个估计量都具有这种形式. 利用 Lagrange 乘法求下式的最小值

$$\sum_{j=1}^{n} a_j^2 \mathrm{Var}(X_j) + \lambda\left(\sum_{j=1}^{n} a_j - 1\right).$$

关于 a_i 的偏导数为

$$2a_i \mathrm{Var}(X_i) + \lambda,$$

令其等于 0, 得到 $a_i = -\lambda[2\mathrm{Var}(X_i)]^{-1}$. 也就是说, X_i 的权重应当与方差的倒数成比例. 而 $\hat{\mu}_2$ 恰好满足该条件, 故它是所有线性估计量中方差最小的.

至于平方和, 考虑

$$\begin{aligned}
\sum_{j=1}^{n} m_j(X_j-\bar{X})^2 &= \sum_{j=1}^{n} m_j(X_j-\mu+\mu-\bar{X})^2 \\
&= \sum_{j=1}^{n} m_j(X_j-\mu)^2 + 2\sum_{j=1}^{n} m_j(X_j-\mu)(\mu-\bar{X}) + \sum_{j=1}^{n} m_j(\mu-\bar{X})^2 \\
&= \sum_{j=1}^{n} m_j(X_j-\mu)^2 + 2(\mu-\bar{X})\sum_{j=1}^{n} m_j(X_j-\mu) + m(\mu-\bar{X})^2 \\
&= \sum_{j=1}^{n} m_j(X_j-\mu)^2 + 2(\mu-\bar{X})m(\bar{X}-\mu) + m(\mu-\bar{X})^2 \\
&= \sum_{j=1}^{n} m_j(X_j-\mu)^2 - m(\bar{X}-\mu)^2. \qquad (16.11)
\end{aligned}$$

对两边取期望得

$$\begin{aligned}
\mathrm{E}\left[\sum_{j=1}^{n} m_j(X_j-\bar{X})^2\right] &= \sum_{j=1}^{n} m_j \mathrm{E}[(X_j-\mu)^2] - m\mathrm{E}[(\bar{X}-\mu)^2] \\
&= \sum_{j=1}^{n} m_j \mathrm{Var}(X_j) - m\mathrm{Var}(\bar{X}) \\
&= \sum_{j=1}^{n} m_j\left(\beta+\frac{\alpha}{m_j}\right) - \beta\left(m^{-1}\sum_{j=1}^{n} m_j^2\right) - \alpha,
\end{aligned}$$

因此

$$\mathrm{E}\left[\sum_{j=1}^{n} m_j(X_j-\bar{X})^2\right]=\beta\left(m-m^{-1}\sum_{j=1}^{n} m_j^2\right)+\alpha(n-1). \tag{16.12}$$

它提供了一个比 (16.9) 式更具一般性的无偏估计量. 令 $\alpha=0$, $m_j=1$, $j=1,2,\cdots n$, 则有 $m=n$. 另外, 若令 $\beta=0$, 从 (16.12) 式可以得到关于 α 的估计量, 其中 X_j 可看作 m_j 个独立观测的平均, 每个观测的均值是 μ、方差是 α. 通常认为诸 m_j 和 m 是已知的. □

习题

16.1 设 X 是参数为 n_1 和 p 的二项分布, 即

$$f_X(x)=\binom{n_1}{x}p^x(1-p)^{n_1-x},\quad x=0,1,2,\cdots,n_1.$$

同时设 Z 是与 X 独立的参数为 n_2 和 p 的二项分布. 试证明 $Y=X+Z$ 是参数为 n_1+n_2 和 p 的二项分布, 并计算给定 $Y=y$ 时 X 的条件分布.

16.2 设 X 和 Y 的联合概率分布如下:

x	y		
	0	1	2
0	0.20	0	0.10
1	0	0.15	0.25
2	0.05	0.15	0.10

(a) 分别计算 X 和 Y 的边缘分布.

(b)计算 $Y=y$, $y=0,1,2$ 时 X 的条件分布.

(c) 分别计算 $y=0,1,2$ 时 $\mathrm{E}(X|y), E(X^2|y)$ 以及 $\mathrm{Var}(X|y)$ 的值.

(d) 用 (c)、(16.5) 式和 (16.8) 式计算 $\mathrm{E}(X)$ 和 $\mathrm{Var}(X)$.

16.3 设 X 和 Y 是两个随机变量, 满足二维正态分布, 联合密度函数为

$$f_{X,Y}(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\times\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2-2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right)+\left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]\right\}.$$

证明 (a) 条件密度函数为

$$f_{X|Y}(x|y)=\frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2}\left[\frac{x-\mu_1-\rho\frac{\sigma_1}{\sigma_2}(y-\mu_2)}{\sigma_1\sqrt{1-\rho^2}}\right]^2\right\}.$$

因此有

$$\mathrm{E}(X|Y=y)=\mu_1+\rho\frac{\sigma_1}{\sigma_2}(y-\mu_2).$$

(b) 边缘概率密度函数为

$$f_X(x)=\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left[-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right].$$

(c) X 和 Y 独立当且仅当 $\rho=0$.

16.4 假设随机变量 $Y_1,\cdots,Y_n$ 相互独立, 且有 $\mathrm{E}(Y_j)=\gamma$ 和 $\mathrm{Var}(Y_j)=a_j+\sigma^2/b_j$, $j=1,2,\cdots n$. 定义 $b=b_1+b_2+\cdots+b_n$ 和 $\bar{Y}=\sum\limits_{j=1}^{n}\frac{b_j}{b}Y_j$. 证明

$$\mathrm{E}\left[\sum_{j=1}^{n}b_j(Y_j-\bar{Y})^2\right]=(n-1)\sigma^2+\sum_{j=1}^{n}a_j\left(b_j-\frac{b_j^2}{b}\right).$$

16.5 给定 $\Theta=(\Theta_1,\Theta_2)$ 的条件下, 随机变量 X 服从均值为 Θ_1、方差为 Θ_2 的正态分布.

(a) 证明 $\mathrm{E}(X)=\mathrm{E}(\Theta_1)$ 和 $\mathrm{Var}(X)=\mathrm{E}(\Theta_2)+\mathrm{Var}(\Theta_1)$.

(b) 如果 Θ_1 和 Θ_2 独立, 证明 X 与 Θ_1+Y 同分布, 其中 Θ_1 和 Y 独立, 且 Y 关于 Θ_2 的条件分布是均值为 0, 方差为 Θ_2 的正态分布.

16.6 假设 Θ 的概率密度函数为 $\pi(\theta),\theta>0$, 且 Θ_1 的概率密度函数为 $\pi_1(\theta)=\pi(\theta-\alpha)$, $\theta>\alpha>0$. 如果在给定 Θ_1 的条件下, X 服从均值为 Θ_1 的 Poisson 分布, 证明 X 与 $Y+Z$ 同分布, 其中 Y 和 Z 独立, Y 服从均值为 α 的 Poisson 分布, $Z|\Theta$ 服从均值为 Θ 的 Poisson 分布.

16.3 有限波动信度理论

信度理论的这一分支是量化可信度问题的第一次尝试. 该方法于二十世纪早期提出, 当时与工伤保险相联系. 最初出现在 1914 年 Mowbray[96]. 问题的提出如下：假设某投保人在过去的第 j 个时期发生了 X_j 的索赔次数或损失额[①], $j\in\{1,2,3,\cdots,n\}$. 也可将 X_j 看成是一列经验数据, 来自保单组中第 j 张保单或是费率系统中某一类的第 j 个成员. 假设 $\mathrm{E}(X_j)=\xi$, 即期望不随时间或成员的不同而改变[②]. 如果能够计算这个值, 那么它就是净保费的数额 (除去各种相关费用, 利润, 还有应对突发情况的准备金). 还假设 $\mathrm{Var}(X_j)=\sigma^2$ 对任意 j 均成立. 用 $\bar{X}=n^{-1}(X_1+\cdots+X_n)$ 来概括所有的历史数据. 注意到 $\mathrm{E}(\bar{X})=\xi$, 若 X_j 相互

① "索赔次数" 指的是索赔发生的次数, 而 "损失额" 指的是支付的数额. 在许多情形如这里的介绍可以指两者任一.

② 这里并没有用常见的符号 μ 表示均值, 因为将在 16.4 节中用 μ 来表示另一个与之相关的均值. 并且, 选取符号 ("Xi") 不但因为它是希腊字母中最难拼写和发音的, 还因为有如下不成文的规定: 在教科书的写作中至少要出现一次这个符号.

独立, 则还有 $\mathrm{Var}(\bar{X})=\sigma^2/n$. 保险人的目标是确定 ξ. 一种做法是忽略过去的数据 (认为其没有可信度), 由其他类似投保人的经验得到某个数 M, 并以它作为保费. 通常称这种保费为**手册费率**, 因为常常用费率手册表示. 另一种做法是不考虑 M 而直接按 $\bar{X}$ 收取保费 (完全信度). 第三种做法是综合考虑 M 和 $\bar{X}$(部分信度).

从保险人的角度来看, 如果过去的经验比较 "稳定"(即变化幅度小, σ^2 较小), 那么就应该更 "偏向于" $\bar{X}$, 也就是说数据 $\bar{X}$ 对下一年保费的预测更有价值. 反之, 如果过去的数据不够稳定, 则 $\bar{X}$ 用于预测次年保费的价值就变小, 这时选用 M 就更为合理.

还有, 如果事先已经知道某投保人很可能会和手册费率中相对应的投保人群体有明显的不同, 则应当赋予 $\bar{X}$ 更多的权重, 因为 $\bar{X}$ 作为一个无偏估计量能表现出关于 ξ 的一些有用信息, 相比之下 M 很可能没有使用价值. 反之, 若其他所有投保人有相近的 ξ 值, 则 M 很可能已经给出了关于索赔次数或损失量的较好描述, 于是就没有必要依靠任何一个个体的历史数据 (可能这些数据还不够完整) 了.

对投保人而言, X_j 可以代表单个投保人、一类拥有相似承保特征的投保人, 或是有某些相似性的团体的相关数据. 例如, 对特定的年份 j, X_j 可以是一张一年期汽车险保单的索赔次数, 可以是某个费率级别 (比如居住在城区的 25 岁以下单身男性, 每年行驶里程超过 7 500 英里) 中所有投保人的平均索赔次数, 也可以是某个食品批发商的运输车队每辆车的平均损失额.

下面首先提供一种判断是否赋予完全信度 (即收取 $\bar{X}$) 的方法, 然后再说明当完全信度条件不满足时, 如何确定部分信度.

16.3.1 完全信度

一种量化 $\bar{X}$ 的稳定性的准则是, 如果有较高的概率, 使得 $\bar{X}$ 和 ξ 的差相对于 ξ 而言较小, 则认为 $\bar{X}$ 是稳定的. 用统计学的术语表示, 就是有两个数 $r>0$ 和 $0<p<1$(r 接近 0, 而 p 接近于 1, 常选取 $r=0.05, p=0.9$) 满足

$$\Pr(-r\xi \leqslant \bar{X}-\xi \leqslant r\xi) \geqslant p \tag{16.13}$$

时, 则赋予 $\bar{X}$ 完全信度.

将 (16.13) 式变形可得

$$\Pr\left(\left|\frac{\bar{X}-\xi}{\sigma/\sqrt{n}}\right| \leqslant \frac{r\xi\sqrt{n}}{\sigma}\right) \geqslant p.$$

定义 y_p 为

$$y_p=\inf_y\left\{\Pr\left(\left|\frac{\bar{X}-\xi}{\sigma/\sqrt{n}}\right| \leqslant y\right) \geqslant p\right\}. \tag{16.14}$$

即 y_p 是满足 (16.14) 式括号中概率条件的最小值. 如果 $\bar{X}$ 有连续的分布函数, 则 (16.14) 式中的 "$\geqslant$" 号可以换成 "$=$" 号, 且 y_p 满足

$$\Pr\left(\left|\frac{\bar{X}-\xi}{\sigma/\sqrt{n}}\right| \leqslant y_p\right) = p. \tag{16.15}$$

此时符合完全信度的条件是 $r\xi\sqrt{n}/\sigma \geqslant y_p$,

$$\frac{\sigma}{\xi} \leqslant \frac{r}{y_p}\sqrt{n} = \sqrt{\frac{n}{\lambda_0}}, \tag{16.16}$$

其中 $\lambda_0 = (y_p/r)^2$. 条件 (16.16) 指出, 当变异系数 σ/ξ 不大于 $\sqrt{n/\lambda_0}$ 时赋予完全信度, 这也与直觉上的判断相一致.

值得注意的是 (16.16) 式可以改写为

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n} \leqslant \frac{\xi^2}{\lambda_0}. \tag{16.17}$$

另外, 在 (16.16) 式中解出 n, 得到符合完全信度条件时, 风险量 (保单数) 应满足

$$n \geqslant \lambda_0 \left(\frac{\sigma}{\xi}\right)^2. \tag{16.18}$$

在许多情形下, 用均值为 ξ, 方差为 σ^2/n 的正态分布可以合理地近似 $\bar{X}$. 比如说, 当 n 足够大时, 可以使用中心极限定理. 这时 $(\bar{X}-\xi)/(\sigma/\sqrt{n})$ 服从标准正态分布, (16.15) 式变成 (Z 服从标准正态分布, $\Phi(y)$ 表示其累积分布函数)

$$\begin{aligned} p &= \Pr(|Z| \leqslant y_p) = \Pr(-y_p \leqslant Z \leqslant y_p) = \Phi(y_p) - \Phi(-y_p) \\ &= \Phi(y_p) - 1 + \Phi(y_p) = 2\Phi(y_p) - 1. \end{aligned}$$

从而 $\Phi(y_p) = (1+p)/2$, 故 y_p 是标准正态分布的 $(1+p)/2$ 分位数.

例如, $p = 0.9$, 查标准正态分布表可知 $y_{0.9} = 1.645$. 再令 $r = 0.05$, 则 $\lambda_0 = (32.9)^2 = 1\ 082.41$, 并且从 (16.18) 式可得 $n \geqslant 1\ 082.41\sigma^2/\xi^2$. 这个不等式是建立在 X_j 的变异系数已知的前提下. 就算不知道 ξ 的值 (注意, 它就是要估计的对象), 仍有可能得到关于变异系数的一些信息.

在使用 (16.18) 式的过程中值得注意的是, 变异系数是作为被估计的风险量的一个估计量存在的. 不等式的右边给出了在符合完全信度的情形下, 风险量应满足的条件. 如果要增加风险量, 只需把式子两边同时乘以某个合适的量. 最后, 任何未知量都必须由数据估计, 这意味着信度理论的问题可以有多种提法. 下面的例子覆盖了大部分常见的情形.

例 16.8 假设对某个特定的投保人其过去的损失量 $X_1, \cdots, X_n$. 样本均值将用于估计 $\xi = \mathrm{E}(X_j)$. 确定满足完全信度的标准. 然后设有 10 个观察值, 6 个是 0, 其余的是 253, 398, 439 和 756. 在 $r = 0.05, p = 0.9$ 的条件下, 确定满足完全信度的条件.

解 直接利用 (16.18)

$$n \geqslant \lambda_0 \left(\frac{\sigma}{\xi}\right)^2.$$

对这个特例, 估计均值和方差为 184.6 和 267.89(方差的估计使用了 $n-1$ 个无偏估计量). 又 $\lambda_0 = 1\ 082.41$, 故条件为

$$n \geqslant 1082.41 \left(\frac{267.89}{184.6}\right)^2 = 2\ 279.51.$$

显然 10 个观察值是不满足完全信度的. □

在下面的例子中进一步假设观察值服从某些特定分布.

例 16.9 假设某投保人可用的过去损失量为 $X_1, \cdots, X_n$, 并合理地假设诸 X_j 相互独立, 服从复合 Poisson 分布. 也就是说, $X_j = Y_{j1} + \cdots + Y_{jN_j}$, 且 N_j 服从参数为 λ 的 Poisson 分布. 单个损失分布 Y 的均值是 θ_Y, 方差是 $\sigma_Y{}^2$. 分别就每份保单的索赔次数的期望估计和考虑每份保单索赔额的期望估计, 确定相应的满足完全信度的条件. 然后确定例 16.8 是否满足条件, 其中已知前三笔非零赔付分别代表单个赔案, 而最后一笔非零赔付由两个赔案 129 和 627 组成.

解 **情况 1** 考虑平均索赔次数. 使用 N_j 而非 X_j, 有 $\xi = \mathrm{E}(N_j) = \lambda$ 和 $\sigma^2 = \mathrm{Var}(N_j) = \lambda$, 由 (16.18) 式得

$$n \geqslant \lambda_0 \left(\frac{\lambda^{1/2}}{\lambda}\right)^2 = \frac{\lambda_0}{\lambda}.$$

因此, 如果用保单数目表示完全信度条件, 则要求保单数超过 λ_0/λ, 其中的 λ 由数据估计. 如果条件是用期望索赔次数来表示, 即 $n\lambda$, 将上式两边同时乘以 λ, 得到

$$n\lambda \geqslant \lambda_0.$$

看上去这个条件似乎不需要做任何估计, 实际上需要估计索赔次数的期望值. 实际运用中, 这个条件是用实际发生的索赔次数来表示, 即把左边的 $n\lambda$ 换为估计量 $N_1 + \cdots + N_n$.

对题设数据来说, 共有 5 次索赔, 故 λ 的估计值是 0.5. 满足完全信度的条件变成

$$n \geqslant \frac{1\ 082.41}{0.5} = 2\ 164.82,$$

从而 10 张保单是远远达不到这个条件的. 也可以将 5 笔实际的赔案和 λ_0=1 082.41 作比较, 结论相同.

情况 2 考虑平均总赔付额, 首先利用第 6 章中的公式, 得到 $\xi = \mathrm{E}(X_j) = \lambda\theta_Y$ 和 $\mathrm{Var}(X_j) = \lambda(\theta_Y^2 + \sigma_Y^2)$. 用样本数量来表示, 条件为

$$n \geqslant \lambda_0 \frac{\lambda(\theta_Y^2 + \sigma_Y^2)}{\lambda^2 \theta_Y^2} = \frac{\lambda_0}{\lambda}\left[1 + \left(\frac{\sigma_Y}{\theta_Y}\right)^2\right].$$

若要将条件用索赔次数的期望来表示, 将上式两边同时乘以 λ, 可得

$$n\lambda \geqslant \lambda_0 \left[1 + \left(\frac{\sigma_Y}{\theta_Y}\right)^2\right].$$

最后, 如果条件是用总的索赔金额的期望来表示, 将上式两边再同时乘以 θ_Y 得到

$$n\lambda\theta_Y \geqslant \lambda_0 \left(\theta_Y + \frac{\sigma_Y^2}{\theta_Y}\right).$$

对题设数据, 5 笔索赔的均值是 369.2, 标准差是 189.315, 因此有

$$n \geqslant \frac{\lambda_0}{\lambda}\left[1 + \left(\frac{\sigma_Y}{\theta_Y}\right)^2\right] = \frac{1\ 082.41}{0.5}\left[1 + \left(\frac{189.315}{369.2}\right)^2\right] = 2\ 734.02.$$

10 个观察值还是远远达不到要求的. 如果用索赔次数表示条件, 则两边同乘 0.5, 得到 1 367.01. 最后, 用总索赔金额来表示条件, 只需将两边同乘 369.2 得到 504 701. 注意到这 3 种情形中, 观测数量和需要达到的数值之比是不变的:

$$\frac{10}{2\ 734.02} = \frac{5}{1\ 367.01} = \frac{1\ 846}{504\ 701} = 0.003\ 658. \qquad \square$$

这些例子并不满足完全信度的条件, 因此仅用样本均值并不能充分准确地估计期望值, 还需要另外的处理办法.

16.3.2 部分信度

如果已经确定不满足完全信度的条件, 那么出于竞争或其他原因, 在净保费中既能够反映历史的经验 $\bar{X}$ 又能够反映外来的均值 M 似乎会更恰当. 直觉的方法是运用加权平均, 也就是说, 用信度保费

$$P_c = Z\bar{X} + (1 - Z)M, \tag{16.19}$$

其中信度因子 $Z \in [0,1]$ 待定. 精算学有许多选取 Z 的方法, 通常都是基于直觉而非理论. (之前提到 Mowbray[96] 考虑的是完全信度, 不是部分信度.) 一种重要的选择方法是

$$Z = \frac{n}{n+k}, \tag{16.20}$$

其中 k 待定. 16.4 节将从理论上利用一个统计模型证明其合理性. 现在要讨论另外一种方法, 它和完全信度的想法相同 (且包含了完全信度的情形, 即 $Z = 1$).

关于 Z 值的选取有各种各样的论证方法, 其中许多种方法导出的结论相同. 所有的方法都或多或少地存在一些缺点. 虽然这里给出的方法也有缺陷, 但至少它是简单明了的. 回忆完全信度条件, 它的目标是保证所考虑的净保费 ($\bar{X}$) 和本来应当使用的量 (ξ) 之间的差以很大的概率达到很小. 由于 $\bar{X}$ 是无偏的, 这在本质上 (如果 $\bar{X}$ 确实服从正态分布) 等价于控制所提出的净保费 $\bar{X}$ 的方差. 从 (16.17) 式可以看出, 虽然无法保证 $\bar{X}$ 的方差足够小, 但是控制信度保费 P_c 的方差却是可行的, 方法如下:

$$\frac{\xi^2}{\lambda_0} = \mathrm{Var}(P_c) = \mathrm{Var}[Z\bar{X} + (1-Z)M] = Z^2\mathrm{Var}(\bar{X}) = Z^2\frac{\sigma^2}{n}.$$

因此如果 Z 比 1 小, 则有 $Z = (\xi/\sigma)\sqrt{n/\lambda_0}$. 可以统一写为

$$Z = \min\left\{\frac{\xi}{\sigma}\sqrt{\frac{n}{\lambda_0}}, 1\right\}. \tag{16.21}$$

(16.21) 式的一个解释是, 信度因子 Z 是完全信度条件所要求的变异系数 ($\sqrt{n/\lambda_0}$) 与实际变异系数的比值. 常称这种方法为部分信度的平方根法则.

还可以对 (16.21) 式进行代数变形, 不过只需要记住, Z 就是实际保单数与完全信度所要求的保单数之比的平方根.

例 16.10 假设例 16.8 中的手册费率 M 为 225, 求信度保费的估计.

解 平均赔付是 184.6. 利用平方根法则, 信度因子是

$$Z = \sqrt{\frac{10}{2\ 279.51}} = 0.066\ 23.$$

因此信度保费是

$$P_c = 0.066\ 23(184.6) + 0.933\ 77(225) = 222.32.$$

□

例 16.11 假设例 16.9 中的手册费率 M 为 225, 求两种情形下信度保费的估计.

解 第一种情形, 信度因子是

$$Z = \sqrt{\frac{5}{1\ 082.41}} = 0.067\ 97,$$

进而可得

$$P_c = 0.067\ 97(184.6) + 0.932\ 03(225) = 222.25.$$

初看起来似乎不太妥当. 完全信度的条件是以频率估计的形式表示的, 而却用于估计总索赔额. 但通常的情形是, 个体之间的区别更多地体现于索赔频率上, 而不是每笔索赔额的差别上. 所以上面的因子可以体现大部分重要特征.

至于第二种情形, 可以采用以下任何一种计算方法:

$$Z=\sqrt{\frac{10}{2\ 734.02}}=\sqrt{\frac{5}{1\ 367.01}}=\sqrt{\frac{1\ 846}{504\ 701}}=0.060\ 48.$$

从而有

$$P_c=0.060\ 48(184.6)+0.939\ 52(225)=222.56.$$

□

前面提到, 这种方法是有缺陷的. 其实, 除了假设方差充分地代表了 $\bar{X}$ 的波动性值得商榷以外, 其余数学推导都是正确的. 关键的缺陷在于目标的选择. 与 $\bar{X}$ 不同, P_c 并不是 ξ 的无偏估计量. 事实上, 信度理论能够派上用场的原因之一是使用了有偏估计量, 但那也意味着评价 P_c 好坏的标准将不是它的方差, 而是它的均方误差. 然而, 这又需要知道 ξ 和 M 之间的关系. 然而我们不知道它们之间的任何信息, 收集来的数据也起不了多少作用. 正如在后面章节中提到的, 这不仅仅是如何确定 Z 的问题, 同时也是有限波动方法的本质缺点, 16.4 节将介绍有关这种关系的一个模型.

最后用下面几个例子结束本节. 在前两个例子中均采用 $\lambda_0=1\ 082.41$.

例 16.12 牙医团体保险的过去类似经验表明单个个体每年平均损失的均值是 175, 标准差是 140. 某特殊团体在两年中的第一年有 100 人投保, 第二年中有 110 人, 且在这两年内平均每年的赔付是 150. 首先确定是完全信度还是部分信度, 然后计算若下一年有 125 人投保, 应当收取多少信度保费.

解 下面以个体投保人为单位应用有关结论. 已经观察了 100+110=210 个风险单位, (假设索赔经历与个体和年份无关), 且 $\bar{X}$=150. 现知 $M=175$, 并假设这个群体的 σ 是 140. 由于要估计每人的平均损失, 可以应用例 16.11 的情形 2. 也就是, 已知 $n=210$ 和 $\lambda_0=1\ 082.41$, 用样本均值 150 来估计 θ_Y, 进而得到满足完全信度的条件

$$n\geqslant 1\ 082.41\left(\frac{140}{150}\right)^2=942.90,$$

从而算得

$$Z=\sqrt{\frac{210}{942.90}}=0.472$$

(注意到 $\bar{X}$ 是 210 个赔付的平均值, 由中心极限定理, 可以用正态分布近似). 因此每个人的净保费是

$$P_c=0.472(150)+0.528(175)=163.2.$$

整个团体的净保费是 125(163.2)=20 400. □

例 16.13 假设某保险责任的信度仅仅依赖于索赔次数. 对一个特定的群体, 观察到 715 次索赔, 计算合适的信度因子. 假设索赔次数服从 Poisson 分布.

解 这是例 16.11 中的情形 1. 满足完全信度的条件是 $n\lambda \geqslant \lambda_0 = 1\ 082.41$, 故有

$$Z = \sqrt{\frac{715}{1\ 082.41}} = 0.813. \qquad \square$$

例 16.14 *某特定群体的历史数据是 $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)^{\mathrm{T}}$, 其中 X_j 是独立同分布的复合 Poisson 随机变量, 每笔赔付服从指数分布. 如果根据索赔次数得到的信度因子是 0.8, 求用总索赔额计算的信度因子.*

解 对于服从 Poisson 分布的索赔次数, 从例 16.9 可知, $Z = 0.8$ 意味着 $\lambda_n/\lambda_0 = (0.8)^2 = 0.64$, 其中 λ_n 是观察到的索赔次数. 对服从指数分布的单笔赔付额有 $\sigma_Y^2 = \theta_Y^2$. 从例 16.9 的第二种情形可知, 满足完全信度的条件用索赔次数的形式表示为

$$n\lambda \geqslant \lambda_0 \left[1 + \left(\frac{\sigma_Y}{\theta_Y}\right)^2\right] = 2\lambda_0.$$

因此有

$$Z = \sqrt{\frac{\lambda_n}{2\lambda_0}} = \sqrt{0.32} = 0.566. \qquad \square$$

16.3.3 关于有限波动信度方法的一些问题

虽然用有限波动方法可以得到简明的解决方案, 但在理论上却不能自圆其说. 首先, 诸 X_j 的分布没有潜在的理论模型支持, 因此没有明确证明为什么形如 (16.19) 式的保费形式是恰当的, 并且更优于 M. 另外, 为什么不直接收集风险同质的投保人的数据来估计 ξ, 进而对所有这些投保人收取相同的费率呢? 出于实用的原因使用 (16.19) 式, 但是没有提出任何模型说明这样做是合适的. 因此, Z 的选择 (进而 P_c 的选择) 完全是主观的.

其次, 即使对特定模型而言 (16.19) 式是恰当的, 但却对如何选择 r 和 p 没有提出任何建议.

最后, 有限波动方法没有考虑 ξ 和 M 之间的差别. 当运用 (16.19) 式时, 本质上是宣称 M 能够精确地代表在没有任何其他信息的条件下某投保人的期望损失, 然而, M 本身常常也是一个估计值, 因此并不够可靠. 因此, 正确的信度问题的提法应该是 "与 M 相比, $\bar{X}$ 的可靠程度增加了多少? " 而不是 "$\bar{X}$ 有多可靠? ".

本章后面部分将介绍一种系统化的建模方法. 它是基于特定投保人的索赔历史数据, 并指出过去数据与预期费率的厘定相关. 还有, 很符合直觉的公式 (16.19) 也是这种方法的结论之一, 通常从形如 (16.20) 的关系式中得到 Z 的值.

16.3.4 备注

Herzog[52] 和 Longley-Cook[87] 论述了有限波动方法. 亦见 Norberg[100].

习题

16.7 某保险公司决定对个人费率厘定确立满足完全信度的条件, 其标准是观察到的总赔付额位于期望总赔付额正负 5%范围以内的概率是 0.95. 索赔频率服从 Poisson 分布, 单次索赔额的概率密度函数为

$$f(x)=\frac{100-x}{5\ 000},\quad 0\leqslant x\leqslant 100.$$

利用正态分布近似, 计算满足完全信度所需的索赔次数的期望值.

16.8 已知某投保人过去的索赔数据是 $X_1,\cdots,X_n$, 其中诸 X_j 为独立同分布的复合 Poisson 随机变量, Poisson 参数为 λ, 单次索赔的数额服从概率密度如下的 gamma 分布

$$f_Y(y)=\frac{y^{\alpha-1}\mathrm{e}^{-y/\beta}}{\Gamma(\alpha)\beta^{\alpha}},\quad y>0.$$

同时还已知

(1) 根据索赔次数得到的信度因子是 0.9.

(2) 单次索赔额的期望为 $\alpha\beta=100$.

(3) 根据总索赔额得到的信度因子是 0.8.

求 α 和 β 的值.

16.9 某投保人的手册费率为每年600. 在表 16-1 中给出了其过去的赔付数据. 判定其满足完全信度还是条件信度, 并利用正态分布近似计算下一年的净保费. 取 $r=0.05, p=0.9$.

表 16-1 习题 16.9 的有关数据

年	1	2	3
索赔量	475	550	400

16.10 假定 X_j 服从复合负二项分布而不是复合 Poisson 分布, 重新计算例 16.9.

16.11* 某投保群体的总索赔次数服从均值为 λ 的 Poisson 分布, 利用正态分布近似计算 λ 的值, 使得所观察到的赔付次数位于 λ 的实际值的正负 3%范围以内的概率是 0.975.

16.12* 某保险公司将根据以往的数据修订费率. 在满足完全信度的条件下, 挑选索赔次数的期望值, 使得观察到的总索赔额位于总索赔额期望值正负 5%以内的概率为 90%. 单次索赔额服从 0 到 200 000 的均匀分布, 且索赔次数服从 Poisson 分布. 最近的数据包含了 1 082 笔赔付, 利用正态分布近似计算信度因子 Z.

16.13* 某保险群体的每笔索赔额的平均值是 1 500, 标准差是 7 500. 假设索赔次数服从 Poisson 分布, 求期望索赔次数的值, 使得总索赔额位于期望总索赔额正负 6%范围以内的概率为 0.90.

16.14* 某保险群体共产生了 6 000 笔索赔和 15 600 000 的总损失额. 事前对总损失额的估计是 16 500 000. 试用有限波动信度理论方法估计群体的总损失额, 以习题 16.13 中确定的完全信度为标准.

16.15* 假设完全信度标准是总索赔次数位于真值正负 5%范围以内的概率为 p, 结果得到总索赔次数是 800. 现将标准变为总损失额位于真值正负 10%范围以内的概率为 p. 已知索赔频率服从 Poisson 分布, 单次损失额的概率密度函数是 $f(x) = 0.000\,2(100 - x)$, $0 < x < 100$. 计算在新标准下满足完全信度所需的期望索赔次数.

16.16* 现有一个对 1 000 笔赔付使用的完全信度标准, 它使得实际的净保费位于期望净保费正负 10%范围以内的概率是 95%. 索赔次数服从 Poisson 分布, 求单笔损失分布的变异系数.

16.17* 某投保群体已知如下信息.

(1) 期望总损失的事先估计是 20 000 000.

(2) 观察到的总损失是 25 000 000.

(3) 观察到的索赔次数是 10 000.

(4) 满足完全信度所需的索赔次数是 17 500.

根据以上所有信息, 求该群体的期望总损失的信度估计. 注意在估计期望索赔数时应采用相应的信度因子.

16.18* 现有一个完全信度的标准, 使得总索赔次数位于其期望值正负 5%范围以内的概率为 98%. 如果相同的期望索赔次数用于满足总损失额的完全信度条件的话, 则实际总损失额与期望总损失额之差小于 $100K$%的概率为 95%. 单笔索赔的概率密度函数是 $f(x) = 2.5x^{-3.5}, x > 1$, 索赔次数服从 Poisson 分布. 求 K.

16.19* 已知索赔次数服从 Poisson 分布, 且索赔次数和单笔索赔额独立. 单笔索赔分别以 0.5, 0.3, 0.2 的概率取值 1, 2, 10, 求所需的索赔次数期望值, 使得总索赔额位于期望总索赔额正负 10%范围以内的概率是 90%.

16.20* 已知索赔次数服从 Poisson 分布, 且单笔赔付的变异系数是 2. 估计总损失的完全信度的标准是 3 145, 若用此标准, 则所观察到的净保费位于期望净保费正负 k%范围以内的概率是 95%, 求 k.

16.21* 给定如下条件:

(1) $P=$ 某特殊险种净保费的先验估计.

(2) $O=$ 同一险种最新经验数据计算的净保费.

(3) $R=$ 根据观察数据利用信度因子所修订的净保费估计.

(4) $F=$ 满足完全信度条件所需的索赔次数.

试将观察到的索赔次数表示成以上 4 项的函数.

16.4 最大精度信度理论

16.4.1 引言

本节和 16.5 节将考虑一种基于数学模型解决信度问题的方法. 这种称为最大精度信度理论的方法是 Bühlmann 在 1967 年的经典论文 [18] 的延伸, 其中的许多见解在 Whitney[136] 和 Bailey[8] 中也有提及.

下面回到基本问题. 对特定的投保人而言, 观察到单位风险的 n 个赔付数据 $\boldsymbol{X}=(X_1,X_2,\cdots,X_n)^{\mathrm{T}}$. 适用于该投保人的手册费率记为 μ(不再使用符号 M), 但历史经验表明 μ 可能不合适 [$\bar{X}=n^{-1}(X_1+\cdots+X_n)$, 或者说 $\mathrm{E}(X)$, 可能与 μ 存在相当大的差异]. 这就提出了一个问题, 下一年个体风险单位的净保费应该是 μ 还是 $\bar{X}$, 或是两者兼而有之.

保险人需要考虑如下问题: 是该投保人确实与计算 μ 时所假设的条件有不符合之处, 还是仅仅因概率或偶然性导致了 μ 和 $\bar{X}$ 的差别?

虽然明确地回答上述问题是困难的, 但有一点很显然, 没有一个费率系统是完美的. 手册费率 μ 的计算基础是: (a) 评估投保人的核保特征; (b) 将投保人按照特征归类并收取相应的费率. 每个级别应当包含具有相似承保特点的风险. 换句话说, 每个费率级别的风险在核保特征上是同质的. 问题是, 同一级别中的所有风险显然不可能是完全同质的. 不管当初的划分是多么地细致, 同一等级中依然会存在不同质风险 (即相对更低或更高的风险).

因此, 可能某个投保人的条件确实和事先的假定有所不同. 如果这样应该如何确定合适的费率水平呢?

为了继续讨论, 进一步假设在同一费率级别中每个投保人的风险水平可以用一个风险参数 θ 表示 (θ 也有可能是向量形式), 不同的投保人对应于不同的 θ, 这就把风险特征之间的差异进行了量化. 由于所有的核保特征都已经用于分类, 故可以认为 θ 只代表残差, 即影响风险水平但又无法观测的因素. 我们不仅假设 θ 的存在性, 同时还假设它是无法观察的, 所以它的真值永远无法获知.

因为不同的投保人对应不同的 θ, 所以在每个费率级别中必然存在 θ 的概率分布, 其概率函数记为 $\pi(\theta)$. 如果 θ 是标量, 则 θ 的累积分布函数 $\prod(\theta)$ 可以认为是每个费率级别中风险参数 Θ 小于等于 θ 的投保人所占的比例. [用统计术语来说, Θ 是一个随机变量, 分布函数是 $\prod(\theta)=\Pr(\Theta\leqslant\theta)$.] 也就是说, $\prod(\theta)$ 指的是在每个费率级别中随机挑选的一名投保人的风险参数小于或等于 θ 的概率 (为了使其适用于那些新增的投保人, 这里将费率级别的概念稍加推广, 使之包括所有潜在的拥有类似风险的人, 而不论其是否投保).

虽然不知道具体的单个投保人的 θ 值, 但在本节中假定 $\pi(\theta)$ 已知, 即人群的风险特征的结构是已知的. 这个假定可以放宽, 在后续论述中会讨论如何估计 $\pi(\theta)$ 的有关性质, 这是运用该定理所必须的.

人群中存在着不同的风险等级, 投保人的经验数据也随着 θ 的不同而呈现某种规律性的差异. 随机挑选出一个投保人, 可以想象, 其经验数据是通过一个两阶段的过程产生的. 首先, 风险参数 θ 是从概率分布函数 $\pi(\theta)$ 生成的, 接下来, 索赔或损失额将来自给定 θ 下 X 的条件分布 $f_{X|\Theta}(x|\theta)$. 所以说, 经验数据依赖给定风险参数 θ 下的条件分布函数, 随着 θ 的变化会取不同的值. 不同的投保人对应的条

件分布互不相同, 反映了风险参数之间的差别.

例 16.15 考虑机动车保险的某个费率级别, 其中 θ 表示风险参数是 θ 的投保人的期望索赔次数. 为了表示索赔发生率的差别, 假定在该费率级别中 θ 将取相异的值. 相对而言, 表现好的驾驶员是 θ 值较小的, 而表现差的驾驶员则是 θ 值较大的. 为方便起见, 假设每个风险参数为 θ 的投保人的索赔次数服从均值为 θ 的 Poisson 分布, 并假设随机变量 Θ 服从参数为 α 和 β 的 gamma 分布. 已知该费率级别的平均期望索赔次数是 $0.15[\mathrm{E}(\Theta)=0.15]$, 且 95%的投保人的期望索赔次数是在 0.10 和 0.20 之间, 求 α 和 β 的值.

解 对 gamma 分布使用正态分布近似. 已知对正态分布来说, 数据落在与均值的差的绝对值小于或等于两倍标准差的区间内的概率是 95%, 由此可得 Θ 的标准差是 0.025. 从而有 $\mathrm{E}(\Theta)=\alpha\beta=0.15$ 和 $\mathrm{Var}(\Theta)=\alpha\beta^2=(0.025)^2$, 解得 β=1/240, α=36. □

例 16.16 现有两种类型的驾驶员. 表现好的驾驶员占总人数的 75%, 一年中发生 0 次索赔的概率是 0.7, 一次索赔的概率是 0.2, 两次索赔的概率是 0.1. 表现差的驾驶员占总人数的 25%, 一年中发生 0, 1, 2 次索赔的概率分别是 0.5, 0.3, 0.2. 试建立索赔模型, 并说明它是如何与未知风险参数相联系的.

解 当某个驾驶员购买保险时, 保险人并不知道他是好的驾驶员还是差的驾驶员, 因此风险参数 Θ 是两个值中的某一个. 设 $\Theta=G$ 表示好的驾驶员, $\Theta=B$ 表示差的驾驶员, 关于索赔次数 X 的概率模型和风险参数 Θ 在表 16-2 中给出. □

表 16-2 例 16.16 中的概率

x	$\Pr(X=x\|\Theta=G)$	$\Pr(X=x\|\Theta=B)$	θ	$\Pr(\Theta=\theta)$
0	0.7	0.5	G	0.75
1	0.2	0.3	B	0.25
2	0.1	0.2		

例 16.17 已知单笔索赔额服从均值为 $1/\Theta$ 的指数分布, 所有已投保和可能投保的人的 Θ 服从 α=4, β=0.001 的 gamma 分布. 试用数学语言描述这个模型.

解 对索赔额, 有

$$f_{X|\Theta}(x|\theta)=\theta\mathrm{e}^{-\theta x},\quad x,\theta>0,$$

对风险参数, 有

$$\pi_\Theta(\theta)=\frac{\theta^3\mathrm{e}^{-1\,000\theta}1\,000^4}{6},\quad \theta>0.$$ □

16.4.2 贝叶斯方法

继续假设人群中风险特征的分布可以用 $\pi(\theta)$ 表示, 而且参数是 θ 的特定投保人的经验数据来自给定 θ 时索赔或损失的条件分布 $f_{X|\Theta}(x|\theta)$.

现在回到 16.3 节中提出的问题. 即, 对特定投保人已观察到 $\boldsymbol{X}=\boldsymbol{x}$, 其中 $\boldsymbol{X}=(X_1,\cdots,X_n)^{\mathrm{T}}$, $\boldsymbol{x}=(x_1,\cdots,x_n)^{\mathrm{T}}$, 希望能够为 X_{n+1} 确定费率. 假设投保人对应的风险参数是 θ(未知), 投保人不同时期的数据是相互独立的. 用统计术语来说, 就是索赔或损失额 $X_1,\cdots,X_n,X_{n+1}$ 在给定 θ 的条件下相互独立 (虽然未必同分布).

设 X_j 的条件概率函数为

$$f_{X_j|\Theta}(x_j|\theta),\quad j=1,\cdots,n,\quad n+1,$$

注意到若 X_j 在 $\Theta=\theta$ 的条件下同分布, 则 $f_{X_j|\Theta}(x_j|\theta)$ 与 j 无关. 理想的情形是已知给定 $\Theta=\theta$ 的条件下 X_{n+1} 的条件分布, 以便预测同一投保人的未来索赔 X_{n+1}(假设 θ 没有变). 如果已知 θ 就可以使用 $f_{X_{n+1}|\Theta}(x_{n+1}|\theta)$. 然而并不知道 θ 的值, 但我们已知同一投保人的观测 $\boldsymbol{x}$. 显而易见, 下一步应当以 $\boldsymbol{x}$ 为条件. 因此将计算给定 $\boldsymbol{X}=\boldsymbol{x}$ 下 X_{n+1} 的条件分布, 像 12.4 节中定义的那样, 称其为**预测分布**.

给定 $\boldsymbol{X}=\boldsymbol{x}$ 条件下, X_{n+1} 的预测分布与风险分析、管理和决策相关, 并将索赔额的不确定性与风险参数结合起来.

这里将重复 12.4 节中的推导过程. 若 Θ 为离散分布, 则积分号用求和号代替. 因为诸 X_j 在 $\Theta=\theta$ 的条件下相互独立, 所以有

$$f_{\boldsymbol{x},\Theta}(\boldsymbol{x},\theta)=f(x_1,\cdots,x_n|\theta)\pi(\theta)=\left[\prod_{j=1}^{n}f_{X_j|\Theta}(x_j|\theta)\right]\pi(\theta).$$

故 $\boldsymbol{X}$ 的联合分布是对 θ 积分得到的边缘分布, 也就是

$$f_{\boldsymbol{X}}(\boldsymbol{x})=\int\left[\prod_{j=1}^{n}f_{X_j|\Theta}(x_j|\theta)\right]\pi(\theta)\mathrm{d}\theta.\tag{16.22}$$

类似地, 在 (16.22) 式右边用 $n+1$ 代替 n, 即得到 $X_1,\cdots,X_n,X_{n+1}$ 的联合分布. 最后, 在 $\boldsymbol{X}=\boldsymbol{x}$ 下 X_{n+1} 的条件密度是 $(X_1,\cdots,X_n,X_{n+1})$ 的联合密度除以 X 的联合密度, 即

$$f_{X_{n+1}|\boldsymbol{X}}(x_{n+1}|\boldsymbol{x})=\frac{1}{f_{\boldsymbol{X}}(\boldsymbol{x})}\int\left[\prod_{j=1}^{n+1}f_{X_j|\Theta}(x_j|\theta)\right]\pi(\theta)\mathrm{d}\theta.\tag{16.23}$$

在 (16.23) 式中隐含着一个数学结构. 给定 $\boldsymbol{X}$, Θ 的后验概率密度为

$$\pi_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})=\frac{f_{\boldsymbol{x},\Theta}(\boldsymbol{x},\theta)}{f_{\boldsymbol{X}}(\boldsymbol{x})}=\frac{1}{f_{\boldsymbol{X}}(\boldsymbol{x})}\left[\prod_{j=1}^{n}f_{X_j|\Theta}(x_j|\theta)\right]\pi(\theta).\tag{16.24}$$

也就是说, $\left[\prod_{j=1}^{n} f_{X_j|\Theta}(x_j|\theta)\right]\pi(\theta)=\pi_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})f_{\boldsymbol{X}}(\boldsymbol{x})$, 代入 (16.23) 式可得

$$f_{X_{n+1}|\boldsymbol{X}}(x_{n+1}|\boldsymbol{x})=\int f_{X_{n+1}|\Theta}(x_{n+1}|\theta)\pi_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})\mathrm{d}\theta. \tag{16.25}$$

等式 (16.25) 表明, 给定 $\boldsymbol{X}$ 下 X_{n+1} 的条件分布可以看作是一种混合分布, 其中含有后验分布 $\pi_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$.

后验分布综合了 θ 的先验分布和条件概率的信息, 这在 (16.25) 中已有所体现. 在定理 12.49 中曾提到, 当概率分布属于线性指数分布族并且 $\pi(\theta)$ 是先验分布的共轭分布时, 后验分布有较为简捷的表达式. 这时计算给定 $\boldsymbol{X}$ 时 X_{n+1} 的条件分布将变得简单.

例 16.18 (例 16.16 续) 对某投保人观察到 $x_1=0, x_2=1$. 计算预测分布 $X_3|X_1=0, X_2=1$ 和后验分布 $\Theta|X_1=0, X_2=1$.

解 由 (16.22) 式, 边缘概率为

$$\begin{aligned}f_{\boldsymbol{x}}(0,1)&=\sum_{\theta} f_{X_1|\Theta}(0|\theta)f_{X_2|\Theta}(1|\theta)\pi(\theta)\\&=0.7(0.2)(0.75)+0.5(0.3)(0.25)=0.142\ 5.\end{aligned}$$

同理, 所有三个变量的联合概率为

$$f_{\boldsymbol{x},X_3}(0,1,x_3)=\sum_{\theta} f_{X_1|\Theta}(0|\theta)f_{X_2|\Theta}(1,\theta)f_{X_3|\Theta}(x_3|\theta)\pi(\theta).$$

因而有

$$\begin{aligned}f_{\boldsymbol{x},X_3}(0,1,0)&=0.7(0.2)(0.7)(0.75)+0.5(0.3)(0.5)(0.25)=0.092\ 25,\\f_{\boldsymbol{x},X_3}(0,1,1)&=0.7(0.2)(0.2)(0.75)+0.5(0.3)(0.3)(0.25)=0.032\ 25,\\f_{\boldsymbol{x},X_3}(0,1,2)&=0.7(0.2)(0.1)(0.75)+0.5(0.3)(0.2)(0.25)=0.018\ 00.\end{aligned}$$

故预测分布为

$$f_{X_3,\boldsymbol{x}}(0|0,1)=\frac{0.092\ 25}{0.142\ 5}=0.647\ 368,$$

$$f_{X_3,\boldsymbol{x}}(1|0,1)=\frac{0.032\ 25}{0.142\ 5}=0.226\ 316,$$

$$f_{X_3,\boldsymbol{x}}(2|0,1)=\frac{0.018\ 00}{0.142\ 5}=0.126\ 316,$$

由 (16.24) 式可得后验概率

$$\pi(G|0,1)=\frac{f(0|G)f(1|G)\pi(G)}{f(0,1)}=\frac{0.7(0.2)(0.75)}{0.142\ 5}=0.736\ 842,$$

$$\pi(B|0,1)=\frac{f(0|B)f(1|B)\pi(B)}{f(0,1)}=\frac{0.5(0.3)(0.25)}{0.142\ 5}=0.263\ 158.$$

在不引起歧义的条件下，后面的讨论将略去 f 和 π 的下标. 由 (16.25) 同样可以得到预测概率，从算法的角度来说常常更简单. 即

$$f(0|0,1)=\sum_{\theta} f(0|\theta)\pi(\theta|0,1)=0.7(0.736\ 842)+0.5(0.263\ 158)=0.647\ 368,$$

$$f(1|0,1)=0.2(0.736\ 842)+0.3(0.263\ 158)=0.226\ 316,$$

$$f(2|0,1)=0.1(0.736\ 842)+0.2(0.263\ 158)=0.126\ 316,$$

结果和前面计算相同. □

例 16.19 (例 16.17 续) **现有某投保人的索赔数据 100, 950, 450. 求第 4 次索赔的预测分布以及 Θ 的后验分布.**

解 观察值的边缘密度为

$$\begin{aligned} f(100,950,450) &= \int_0^\infty \theta \mathrm{e}^{-100\theta}\theta \mathrm{e}^{-950\theta}\theta \mathrm{e}^{-450\theta}\frac{1\ 000^4}{6}\theta^3\mathrm{e}^{-1\ 000\theta}\mathrm{d}\theta \\ &= \frac{1\ 000^4}{6}\int_0^\infty \theta^6\mathrm{e}^{-2\ 500\theta}\mathrm{d}\theta = \frac{1\ 000^4}{6}\frac{720}{2\ 500^7}. \end{aligned}$$

类似有

$$\begin{aligned} f(100,950,450,x_4) &= \int_0^\infty \theta \mathrm{e}^{-100\theta}\theta \mathrm{e}^{-950\theta}\theta \mathrm{e}^{-450\theta}\theta^{-\theta x_4}\frac{1\ 000^4}{6}\theta^3\mathrm{e}^{-1\ 000\theta}\mathrm{d}\theta \\ &= \frac{1\ 000^4}{6}\int_0^\infty \theta^7\mathrm{e}^{-(2\ 500+x_4)\theta}\mathrm{d}\theta = \frac{1\ 000^4}{6}\frac{5\ 040}{(2\ 500+x_4)^8}. \end{aligned}$$

则预测概率密度为

$$f(x_4|100,950,450)=\frac{\frac{1\ 000^4}{6}\frac{5\ 040}{(2\ 500+x_4)^8}}{\frac{1\ 000^4}{6}\frac{720}{2\ 500^7}}=\frac{7(2\ 500)^7}{(2\ 500+x_4)^8},$$

是参数为 7 和 2 500 的 Pareto 密度.

下面用一种简便方法计算后验分布. 由于表达式的分母是积分形式，计算结果是一个常数，先暂时不予考虑. 分子可以写为

$$\pi(\theta|100,950,450)\propto \theta \mathrm{e}^{-100\theta}\theta \mathrm{e}^{-950\theta}\theta \mathrm{e}^{-450\theta}\frac{1\ 000^4}{6}\theta^3\mathrm{e}^{-1\ 000\theta},$$

在计算边缘密度时要对以上这项做积分. 因为已经忽略分母中的常数，所以顺势把分子中的常数系数也忽略，只保留含变量 (θ) 的乘积项，得到

$$\pi(\theta|100,950,450)\propto \theta^6\mathrm{e}^{-2\ 500\theta}.$$

对上式求积分，可以得出使其成为概率密度函数所需的常数 (积分值为 1). 注意到这个函数具有参数为 7 和 1/2500 的 gamma 分布的形式，因此有

$$\pi(\theta|100,950,450)=\frac{\theta^6\mathrm{e}^{-2\ 500\theta}2\ 500^7}{\Gamma(7)}.$$

从而可求得如下的预测概率密度

$$\begin{aligned} f(x_4|100,950,450) &= \int_0^\infty \theta \mathrm{e}^{-\theta x_4} \frac{\theta^6 \mathrm{e}^{-2\,500\theta} 2\,500^7}{\Gamma(7)} \mathrm{d}\theta \\ &= \frac{2\,500^7}{6!} \int_0^\infty \theta^7 \mathrm{e}^{-(2\,500+x_4)\theta} \mathrm{d}\theta \\ &= \frac{2\,500^7}{6!} \frac{7!}{(2\,500+x_4)^8}, \end{aligned}$$

结果与前面的相同. □

注意到后验分布和先验分布都是 gamma 分布. 共轭先验分布的概念在 12.4.3 节中介绍过. 这里意味着 $X_{n+1}|\boldsymbol{x}$ 是简单混合变量的混合分布, 这为计算 $X_{n+1}|\boldsymbol{x}$ 的密度带来便利. 习题中有相关的例子.

回到最初的问题, 对特定投保人观察到 $\boldsymbol{X}=\boldsymbol{x}$, 并且我们希望预测 X_{n+1}(或期望值). 一种自然的想法是, 如果已知 θ 则直接计算条件期望 (或个体保费)

$$\mu_{n+1}(\theta) = \mathrm{E}(X_{n+1}|\Theta=\theta) = \int x_{n+1} f_{X_{n+1}|\Theta}(x_{n+1}|\theta)\mathrm{d}x_{n+1}. \tag{16.26}$$

若将上式中的 θ 用 Θ 代替, 然后取期望得到

$$\mu_{n+1} = \mathrm{E}(X_{n+1}) = \mathrm{E}[\mathrm{E}(X_{n+1}|\Theta)] = \mathrm{E}[\mu_{n+1}(\Theta)].$$

因此这个净保费或称集体保费是条件期望的期望值, 也是在不知道个体的任何信息时应当对其收取的保费. 这个保费与个体的风险参数 θ 无关, 也不需要利用个体的经验数据 $\boldsymbol{x}$. 在 θ 未知时, 最好的做法是尝试利用经验数据, 这意味着将使用贝叶斯保费 (预测分布的期望值)

$$\mathrm{E}(X_{n+1}|\boldsymbol{X}=\boldsymbol{x}) = \int x_{n+1} f_{X_{n+1}|\boldsymbol{X}}(x_{n+1}|\boldsymbol{x})\mathrm{d}x_{n+1}. \tag{16.27}$$

在算法上更方便的一种形式是

$$\mathrm{E}(X_{n+1}|\boldsymbol{X}=\boldsymbol{x}) = \int \mu_{n+1}(\theta)\pi_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})\mathrm{d}\theta, \tag{16.28}$$

也就是说, 贝叶斯保费是条件期望的期望, 该期望值受后验分布 $\pi_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$ 的影响. 这里提醒读者, 在离散情形积分号要用求和号代替. 为了证明 (16.28) 式, 由 (16.25) 式可得

$$\begin{aligned} \mathrm{E}(X_{n+1}|\boldsymbol{X}=\boldsymbol{x}) &= \int x_{n+1} f_{X_{n+1}|\boldsymbol{X}}(x_{n+1}|\boldsymbol{x})\mathrm{d}x_{n+1} \\ &= \int x_{n+1} \left[\int f_{X_{n+1}|\Theta}(x_{n+1}|\theta)\pi_{\Theta|\boldsymbol{x}}(\theta|\boldsymbol{x})\mathrm{d}\theta\right] \mathrm{d}x_{n+1} \end{aligned}$$

$$= \int \left[\int x_{n+1} f_{X_{n+1}|\Theta}(x_{n+1}|\theta) \mathrm{d}x_{n+1} \right] \pi_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x}) \mathrm{d}\theta$$
$$= \int \mu_{n+1}(\theta) \pi_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x}) \mathrm{d}\theta.$$

例 16.20　(例 16.18 续) 利用 (16.27) 式和 (16.28) 式计算贝叶斯保费.

解　(无法直接观测的) 条件期望为

$$\mu_3(G) = (0)(0.7) + 1(0.2) + 2(0.1) = 0.4,$$
$$\mu_3(B) = (0)(0.5) + 1(0.3) + 2(0.2) = 0.7.$$

如果与例 16.18 一样, 观察到 $x_1 = 0, x_2 = 1$, 则由 (16.27) 式可以直接得到贝叶斯保费

$$\mathrm{E}(X_3|0,1) = 0(0.647\ 368) + 1(0.226\ 316) + 2(0.126\ 316) = 0.478\ 948.$$

则 (无条件) 净保费为

$$\mu_3 = \mathrm{E}(X_3) = \sum_{\theta} \mu_3(\theta)\pi(\theta) = (0.4)(0.75) + (0.7)(0.25) = 0.475.$$

为了对 $x_1 = 0, x_2 = 1$ 验证 (16.28) 式, 利用例 16.18 中得到的后验分布 $\pi(\theta|0,1)$, 有

$$\mathrm{E}(X_3|0,1) = 0.4(0.736\ 842) + 0.7(0.263\ 158) = 0.478\ 947,$$

与前面结果的细微差异是由于四舍五入造成的. 相比之下, 运用 (16.28) 式的第二种方法比直接用 $X_{n+1}|\boldsymbol{X} = \boldsymbol{x}$ 的条件分布计算来得简单. □

结果在意料之中: 基于两个观察值进行修正的保费介于不依靠任何经验数据所得保费 (0.475) 和仅仅依靠经验数据所得保费 (0.5) 之间.

例 16.21　(例 16.19 续) 求贝叶斯保费.

解　由例 16.19 可得 $\mu_4(\theta) = \theta^{-1}$, 进而由 (16.28) 式有

$$\mathrm{E}(X_4|100, 950, 450) = \int_0^{\infty} \theta^{-1} \frac{\theta^6 \mathrm{e}^{-2\ 500\theta} 2\ 500^7}{720} \mathrm{d}\theta$$
$$= \frac{2\ 500^7}{720} \frac{120}{2\ 500^6} = 416.67.$$

这个结果也可以从附录 A 中关于 gamma 分布各阶矩的公式中得到. 由先验分布可以得到

$$\mu = \mathrm{E}(\Theta^{-1}) = \frac{1\ 000}{3} = 333.33,$$

再次发现贝叶斯保费介于先验估计所得保费和仅依靠经验数据所得保费之间 (样本均值是 500).

由 (16.27) 式, 有

$$\mathrm{E}(X_4|100,950,450)=\frac{2\,500}{6}=416.67,$$

此即为预测得到的 Pareto 分布的期望值. □

例 16.22 设样本数为 n, 先验 gamma 分布的参数为 α 和 β, 其中 β 是常用标度参数的倒数, 推广例 16.21 的结果.

解 计算后验分布如下

$$\pi(\theta|\boldsymbol{x})\propto\left(\prod_{j=1}^{n}\theta\mathrm{e}^{-\theta x_j}\right)\frac{\theta^{\alpha-1}\mathrm{e}^{-\beta\theta}\beta^{\alpha}}{\Gamma(\alpha)}\propto\theta^{n+\alpha-1}\mathrm{e}^{-(\Sigma x_j+\beta)\theta}.$$

上式第二个比例式成立是因为后验概率是 θ 的函数, 所以可以忽略所有不含 θ 的乘积项. 这里不必用积分来确定常数系数, 注意到后验分布是第一个参数为 $n+\alpha$, 标度参数为 $(\sum x_j+\beta)^{-1}$ 的 gamma 分布. X_{n+1} 的贝叶斯估计是利用后验分布所算得的 Θ^{-1} 的期望值, 也就是

$$\frac{\Sigma x_j+\beta}{n+\alpha-1}=\frac{n}{n+\alpha-1}\bar{x}+\frac{\alpha-1}{n+\alpha-1}\frac{\beta}{\alpha-1}.$$

该估计值是观察值和无条件期望值的加权平均, 它具有 (16.19) 式中信度加权的形式. □

随机变量也未必是同分布的, 下面就是一个例子.

例 16.23 假设某群体的风险参数为 θ(未知), 在第 j 年有 m_j 人投保. 记该年度总索赔次数为 N_j, 服从均值为 $m_j\theta$ 的 Poisson 分布, 即对 $j=1,\cdots,n$, 有

$$\Pr(N_j=x|\Theta=\theta)=\frac{(m_j\theta)^x\mathrm{e}^{-m_j\theta}}{x!},\quad x=0,1,2,\cdots.$$

以上是基于如下假设成立的: 每个投保人的索赔次数相互独立, 并且都服从均值为 θ 的 Poisson 分布. 对第 $n+1$ 年投保的 m_{n+1} 人, 用贝叶斯算法求期望索赔次数.

解 由题设条件, 第 j 年每人平均索赔次数为

$$X_j=\frac{N_j}{m_j},\quad j=1,\cdots,n.$$

因此有

$$f_{X_j|\Theta}(x_j|\theta)=\Pr[N_j=m_jx_j|\Theta=\theta].$$

设 Θ 服从参数为 α 和 β 的 gamma 分布, 即

$$\pi(\theta)=\frac{\theta^{\alpha-1}\mathrm{e}^{-\theta/\beta}}{\Gamma(\alpha)\beta^{\alpha}},\quad\theta>0;$$

则后验分布 $\pi_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$ 作为 θ 的函数与

$$\left[\prod_{j=1}^{n} f_{X_j|\Theta}(x_j|\theta)\right]\pi(\theta),$$

成比例. 而上式又与

$$\left[\prod_{j=1}^{n} \theta^{m_j x_j}\mathrm{e}^{-m_j\theta}\right]\theta^{\alpha-1}\mathrm{e}^{-\theta/\beta}=\theta^{\alpha+\sum\limits_{j=1}^{n} m_j x_j-1}\mathrm{e}^{-\theta\left(\beta^{-1}+\sum\limits_{j=1}^{n} m_j\right)}.$$

成比例, 也就是与参数为 $\alpha_*=\alpha+\sum\limits_{j=1}^{n} m_j x_j$ 和 $\beta_*=\left(1/\beta+\sum\limits_{j=1}^{n} m_j\right)^{-1}$ 的 gamma 分布的密度函数成比例. 因此 $\Theta|\boldsymbol{X}$ 也是 gamma 分布, 只是把参数 α,β 换成了 α_*,β_*.

现在可得

$$\mathrm{E}(X_j|\Theta=\theta)=\mathrm{E}\left(\frac{1}{m_j}N_j|\Theta=\theta\right)=\frac{1}{m_j}\mathrm{E}(N_j|\Theta=\theta)=\theta.$$

因此有 $\mu_{n+1}(\theta)=\mathrm{E}(X_{n+1}|\Theta=\theta)=\theta$ 以及 $\mu_{n+1}=\mathrm{E}(X_{n+1})=\mathrm{E}[\mu_{n+1}(\Theta)]=\alpha\beta$, 因为 Θ 是参数为 α 和 β 的 gamma 分布. 利用 (16.28) 式和 $\Theta|\boldsymbol{X}$ 是参数为 α_*,β_* 的 gamma 分布, 可知

$$\begin{aligned}\mathrm{E}(X_{n+1}|\boldsymbol{X}=\boldsymbol{x})&=\int_0^{\infty}\mu_{n+1}(\theta)\pi_{\Theta|\boldsymbol{x}}(\theta|\boldsymbol{x})\mathrm{d}\theta\\&=\mathrm{E}[\mu_{n+1}(\Theta)|\boldsymbol{X}=\boldsymbol{x}]=\mathrm{E}(\Theta|\boldsymbol{X}=\boldsymbol{x})=\alpha_*\beta_*.\end{aligned}$$

定义所观察的个体总数是 $m=\sum\limits_{j=1}^{n} m_j$.

变形可得

$$\mathrm{E}(X_{n+1}|\boldsymbol{X}=\boldsymbol{x})=Z\bar{x}+(1+Z)\mu_{n+1},$$

其中 $Z=m/(m+\beta^{-1})$, $\bar{x}=m^{-1}\sum\limits_{j=1}^{n} m_j x_j$, $\mu_{n+1}=\alpha\beta$. 这里又一次得到了与 (16.19) 式相同形式的表达式.

因此, 下一年的 m_{n+1} 个投保人的贝叶斯期望总索赔数是 $m_{n+1}\mathrm{E}(X_{n+1}|\boldsymbol{X}=\boldsymbol{x})$.

如果令 $m_j=1$, 则得到每年索赔次数是独立同分布的 Poisson 随机变量, 这时 $X_j\equiv N_j$ 对 $j=1,2,\cdots,n$ 成立, 且在给定 θ 的条件下服从均值为 θ 的独立 Poisson 分布. 同时有

$$\mathrm{E}(X_{n+1}|\boldsymbol{X}=\boldsymbol{x})=Z\bar{x}+(1-Z)\mu,$$

其中 $Z=n/(n+\beta^{-1})$, $\bar{x}=n^{-1}\sum\limits_{j=1}^{n}x_j$, 且 $\mu=\alpha\beta$. □

在例 16.22 和例 16.23 中, 贝叶斯估计是样本均值 $\bar{x}$ 和净保费 μ_{n+1} 的加权平均. 从信度理论的角度来看, 这是非常吸引人的. 还有, 信度因子 Z 是关于风险单位数量的增函数, 也就是经验数据越多, Z 就越接近 1, 这与直觉相吻合.

16.4.3 信度保费

在上一节中运用了较为系统的方法来处理投保人的经验数据. 理想情形是, 保险人应当收取个体保费 (条件期望)$\mu_{n+1}(\theta)$, 而不是净保费 $\mu_{n+1}=\mathrm{E}(X_{n+1})$. 由于 θ 无法确定, 所以无法直接收取个体保费, 但是可以用基于经验数据 $\boldsymbol{x}$ 的条件期望来代替, 这就是贝叶斯保费 $\mathrm{E}(X_{n+1}|\boldsymbol{x})$.

然而计算贝叶斯保费可能会遇到困难. 虽然, 在前述例子中计算并不复杂, 但是不能指望这些简单的例子就能概括现实中的保险现象. 不论是利用 (16.27) 式还是 (16.28) 式, 在一些更贴近实际的模型中就能够发现计算 $\mathrm{E}(X_{n+1}|\boldsymbol{x})$ 的困难之处, 可能还要用到数值积分的方法. 也有例外, 如例 16.22 和例 16.23.

下面提供一种由 Bühlmann[18] 在 1967 年提出的替代方法. 回忆本质的问题: 希望使用条件分布 $f_{X_{n+1}|\Theta}(x_{n+1}|\theta)$ 或是条件期望 $\mu_{n+1}(\theta)$ 来估计下一年的赔付. 因为 $\mathbf{x}$ 已经被观察到, 一个想法是用经验数据的线性函数来逼近 $\mu_{n+1}(\theta)$. [实际上, 公式 $Z\bar{X}+(1-Z)\mu$ 就是线性形式.] 因此, 把估计形式固定为 $\alpha_0+\sum\limits_{j=1}^{n}\alpha_jX_j$, 其中 $\alpha_0,\alpha_1,\cdots,\alpha_n$ 待定. 选择系数的目标是使均方误差达到最小, 也就是最小化

$$Q=\mathrm{E}\left\{\left[\mu_{n+1}(\Theta)-\alpha_0-\sum_{j=1}^{n}\alpha_jX_j\right]^2\right\}. \tag{16.29}$$

这里的期望值是基于 $X_1,\cdots,X_n,\Theta$ 的联合分布, 即均方误差是对所有可能的 Θ 和可能的观察值取平均. 利用偏导数来最小化 Q, 有

$$\frac{\partial Q}{\partial\alpha_0}=\mathrm{E}\left\{2\left[\mu_{n+1}(\Theta)-\alpha_0-\sum_{j=1}^{n}\alpha_jX_j\right](-1)\right\}.$$

把让 (16.29) 式取到最小值的 $\alpha_0,\alpha_1,\cdots,\alpha_n$ 记作 $\tilde{\alpha}_0,\tilde{\alpha}_1,\cdots,\tilde{\alpha}_n$. 令 $\partial Q/\partial\alpha_0$ 等于 0 可得

$$\mathrm{E}[\mu_{n+1}(\Theta)]=\tilde{\alpha}_0+\sum_{j=1}^{n}\tilde{\alpha}_j\mathrm{E}(X_j).$$

由于 $\mathrm{E}(X_{n+1})=\mathrm{E}[\mathrm{E}(X_{n+1}|\Theta)]=\mathrm{E}[\mu_{n+1}(\Theta)]$, 所以 $\partial Q/\partial\alpha_0=0$ 意味着

$$\mathrm{E}(X_{n+1})=\tilde{\alpha}_0+\sum_{j=1}^{n}\tilde{\alpha}_j\mathrm{E}(X_j). \tag{16.30}$$

方程 (16.30) 可以称作**无偏性方程**, 因为它要求 $\tilde{\alpha}_0+\sum\limits_{j=1}^{n}\tilde{\alpha}_jX_j$ 是 $\mathrm{E}(X_{n+1})$ 的无偏估计. 注意, 信度保费对要估计的 $\mu_{n+1}(\theta)=\mathrm{E}(X_{n+1}|\theta)$ 而言可能是有偏的, 这个偏差将会在所有 Θ 取值范围内得到平均. 在接受偏差的前提下, 可以减少总的均方误差. 对 $i=1,2,\cdots,n$, 有

$$\frac{\partial Q}{\partial\alpha_i}=\mathrm{E}\left\{2\left[\mu_{n+1}(\Theta)-\alpha_0-\sum_{j=1}^{n}\alpha_jX_j\right](-X_i)\right\}.$$

令上式为 0, 得

$$\mathrm{E}[\mu_{n+1}(\Theta)X_i]=\tilde{\alpha}_0\mathrm{E}(X_i)+\sum_{j=1}^{n}\tilde{\alpha}_j\mathrm{E}(X_iX_j).$$

方程左边可以改写为

$$\begin{aligned}\mathrm{E}[\mu_{n+1}(\Theta)X_i]&=\mathrm{E}\{\mathrm{E}[X_i\mu_{n+1}(\Theta)|\Theta]\}=\mathrm{E}\{\mu_{n+1}(\Theta)\mathrm{E}[X_i|\Theta]\}\\&=\mathrm{E}[\mathrm{E}(X_{n+1}|\Theta)\mathrm{E}(X_i|\Theta)]=\mathrm{E}[\mathrm{E}(X_{n+1}X_i|\Theta)]\\&=\mathrm{E}(X_iX_{n+1}),\end{aligned}$$

其中倒数第二个等号是由于在给定 Θ 下 X_i 和 X_{n+1} 独立. 这样 $\partial Q/\partial\alpha_i=0$ 等价于

$$\mathrm{E}(X_iX_{n+1})=\tilde{\alpha}_0\mathrm{E}(X_i)+\sum_{j=1}^{n}\tilde{\alpha}_j\mathrm{E}(X_iX_j).\tag{16.31}$$

将 (16.30) 式两边乘以 $\mathrm{E}(X_i)$, 再用 (16.31) 去减, 得到

$$\mathrm{Cov}(X_i,X_n+1)=\sum_{j=1}^{n}\tilde{\alpha}_j\ \mathrm{Cov}(X_i,X_j),\quad i=1,\cdots,n.\tag{16.32}$$

方程 (16.30) 与 (16.32) 中的 n 个方程一起组成**正交方程组**. 通过解此方程组得到 $\tilde{\alpha}_0,\tilde{\alpha}_1,\cdots,\tilde{\alpha}_n$ 的值, 进而得到信度保费

$$\tilde{\alpha}_0+\sum_{j=1}^{n}\tilde{\alpha}_jX_j.\tag{16.33}$$

虽然能用矩阵直接表示正交方程组的解 (如果诸 X_j 的协方差矩阵非奇异), 但也只能对一些特殊情形得到明确的解的表达式.

注意到 (16.32) 右边只有一项 $\mathrm{Cov}(X_i,X_i)=\mathrm{Var}(X_i)$ 是方差项, 另外 $n-1$ 项都是协方差项.

另外还可以证明, $\tilde{\alpha}_0, \tilde{\alpha}_1, \cdots, \tilde{\alpha}_n$ 也将使

$$Q_1 = \mathrm{E}\left\{\left[\mathrm{E}(X_{n+1}|\boldsymbol{X}) - \alpha_0 - \sum_{j=1}^{n}\alpha_j X_j\right]^2\right\} \tag{16.34}$$

和

$$Q_2 = \mathrm{E}\left[\left(X_{n+1} - \alpha_0 - \sum_{j=1}^{n}\alpha_j X_j\right)^2\right]. \tag{16.35}$$

取到最小值. 只需将 (16.34) 和 (16.35) 分别对 $\alpha_0, \alpha_1, \cdots, \alpha_n$ 求导, 就能看到使导数为 0 的解也满足正交方程组 (16.30) 和 (16.32). 由此可知, 信度保费 (16.33) 同时是条件期望 $\mathrm{E}(X_{n+1}|\Theta)$, 贝叶斯保费 $\mathrm{E}(X_{n+1}|\boldsymbol{X})$ 和 X_{n+1} 的最佳线性估计.

例 16.24 已知 $\mathrm{E}(X_j) = \mu, \mathrm{Var}(X_j) = \sigma^2$, 且对 $i \neq j, \mathrm{Cov}(X_i, X_j) = \rho\sigma^2$, 其中相关系数 ρ 满足 $-1 < \rho < 1$, 求信度保费 $\tilde{\alpha}_0 + \sum\limits_{j=1}^{n}\tilde{\alpha}_j X_j$.

解 由无偏方程 (16.30) 可得

$$\mu = \tilde{\alpha}_0 + \mu\sum_{j=1}^{n}\tilde{\alpha}_j,$$

即

$$\sum_{j=1}^{n}\tilde{\alpha}_j = 1 - \frac{\tilde{\alpha}_0}{\mu}.$$

(16.32) 式的 n 个方程化为, 对 $i = 1, 2, \cdots, n$,

$$\rho = \sum_{\substack{j=1\\ j\neq i}}^{n}\tilde{\alpha}_j\rho + \tilde{\alpha}_i,$$

或者写为

$$\rho = \sum_{j=1}^{n}\tilde{\alpha}_j\rho + \tilde{\alpha}_i(1-\rho), \quad i = 1, \cdots, n.$$

因此有

$$\tilde{\alpha}_i = \frac{\rho\left(1 - \sum_{j=1}^{n}\tilde{\alpha}_j\right)}{1-\rho} = \frac{\rho\tilde{\alpha}_0}{\mu(1-\rho)},$$

最后一步用到无偏方程. 将 i 从 1 到 n 求和得到

$$\sum_{i=1}^{n}\tilde{\alpha}_i = \sum_{j=1}^{n}\tilde{\alpha}_j = \frac{n\rho\tilde{\alpha}_0}{\mu(1-\rho)},$$

联立无偏方程, 有

$$1-\frac{\tilde{\alpha}_0}{\mu}=\frac{n\rho\tilde{\alpha}_0}{\mu(1-\rho)}.$$

解得

$$\tilde{\alpha}_0=\frac{(1-\rho)\mu}{1-\rho+n\rho}.$$

进而有

$$\tilde{\alpha}_j=\frac{\rho\tilde{\alpha}_0}{\mu(1-\rho)}=\frac{\rho}{1-\rho+n\rho}.$$

于是求得信度保费为

$$\tilde{\alpha}_0+\sum_{j=1}^{n}\tilde{\alpha}_jX_j=\frac{(1-\rho)\mu}{1-\rho+n\rho}+\sum_{j=1}^{n}\frac{\rho X_j}{1-\rho+n\rho}=(1-Z)\mu+Z\bar{X},$$

其中 $Z=n\rho/(1-\rho+n\rho)$ 和 $\bar{X}=n^{-1}\sum\limits_{j=1}^{n}X_j$. 因此, 如果 $0<\rho<1$, 则 $0<Z<1$, 信度保费就是 $\mu=\mathrm{E}(X_{n+1})$ 和 $\bar{X}$ 的加权平均, 它具有 (16.19) 式的形式. □

下面给出一些能够具体计算 $X_j|\Theta$ 的条件均值和方差的模型, 进而可以计算均值 $\mathrm{E}(X_j)$, 方差 $\mathrm{Var}(X_j)$ 以及协方差 $\mathrm{Cov}(X_i,X_j)$.

16.4.4 Bühlmann 模型

这是最早期, 也是最简单的信度模型. 具体假定每个投保人 (给定 Θ 的条件下) 过去的损失 $X_1,\cdots,X_n$ 有相同的均值和方差, 并且在给定 Θ 的条件下独立同分布.

定义

$$\mu(\theta)=\mathrm{E}(X_j|\Theta=\theta),$$

$$v(\theta)=\mathrm{Var}(X_j|\Theta=\theta).$$

如前所述, $\mu(\theta)$ 称为**条件期望**, $v(\theta)$ 称为**条件方差**. 定义

$$\mu=\mathrm{E}[\mu(\Theta)],\tag{16.36}$$

$$v=\mathrm{E}[v(\Theta)],\tag{16.37}$$

$$a=\mathrm{Var}[\mu(\Theta)].\tag{16.38}$$

(16.36) 中的 μ 称为**条件期望的期望**, (16.37) 中的 v 称为**条件方差的期望**, (16.38) 中的 a 称为**条件期望的方差**. 注意 μ 就是在没有任何关于 θ(自然也没有关于 $\mu(\theta)$) 的信息下所用的估计量. μ 也称为**集体保费**.

下面可以计算诸 X_j 的均值, 方差和协方差. 首先有

$$\mathrm{E}(X_j)=\mathrm{E}[\mathrm{E}(X_j|\Theta)]=\mathrm{E}[\mu(\Theta)]=\mu. \tag{16.39}$$

其次有

$$\begin{aligned}\mathrm{Var}(X_j)&=\mathrm{E}[\mathrm{Var}(X_j|\Theta)]+\mathrm{Var}[\mathrm{E}(X_j|\Theta)]\\&=\mathrm{E}[v(\Theta)]+\mathrm{Var}[\mu(\Theta)]=v+a.\end{aligned} \tag{16.40}$$

最后对 $i\neq j$, 有

$$\begin{aligned}\mathrm{Cov}(X_i,X_j)&=\mathrm{E}(X_iX_j)-\mathrm{E}(X_i)\mathrm{E}(X_j)=\mathrm{E}[\mathrm{E}(X_iX_j|\Theta)]-\mu^2\\&=\mathrm{E}[\mathrm{E}(X_i|\Theta)\mathrm{E}(X_j|\Theta)]-\{\mathrm{E}[\mu(\Theta)]\}^2\\&=\mathrm{E}\{[\mu(\Theta)]^2\}-\{\mathrm{E}[\mu(\Theta)]\}^2=\mathrm{Var}[\mu(\Theta)]=a.\end{aligned} \tag{16.41}$$

这与例 16.24 中令变量 $\mu=\mu,\sigma^2=v+a,\rho=a/(v+a)$ 形式相同. 因此得到信度保费

$$\tilde{\alpha}_0+\sum_{j=1}^{n}\tilde{\alpha}_jX_j=Z\bar{X}+(1-Z)\mu, \tag{16.42}$$

其中

$$Z=\frac{n}{n+k}, \tag{16.43}$$

$$k=\frac{v}{a}=\frac{\mathrm{E}[\mathrm{Var}(X_j|\Theta)]}{\mathrm{Var}[\mathrm{E}(X_j|\Theta)]}. \tag{16.44}$$

(16.43) 式中的信度因子 Z 称为 Bühlmann 信度因子. 注意 (16.42) 式与 (16.19) 式形式相同, 而 (16.43) 式就是 (16.20) 式. 从 (16.44) 式已经明确知道如何确定 k.

公式 (16.42) 有许多引人注目的性质. 首先是信度保费 (16.42) 是样本均值 $\bar{X}$ 和集体保费 μ 的加权平均, 这正是我们想要的. 还有, 当 n 增大时 Z 趋近于 1, 即随着经验数据的增加而给予 $\bar{X}$ 而不是 μ 更多的信度, 这也与直觉相吻合. 另外, 如果人群关于风险参数 Θ 确实是风险同质的, 则相对而言, 当 Θ 变动时条件期望 $\mu(\Theta)=\mathrm{E}(X_j|\Theta)$ 不会有太大的变化 (也就是它们的值接近). 因此 a 相对 v 来说会比较小, 使得 k 较大, 从而 Z 接近 0. 这不违反直觉, 因为对风险同质的人群来说, 总的平均值 μ 在预测单个投保人下一年的索赔时将会更有价值. 反之, 对风险差异较大的人群来说, 条件期望 $\mathrm{E}(X_j|\Theta)$ 变动会更大, 也就是 a 较大, k 较小, 因此 Z 接近 1. 这也是合乎情理的, 因为在风险异质的人群中, 预测单独一个投保人的未来损失时, 其他投保人的经历应该比投保人本人的经验数据的价值小.

下面是几个例子.

例 16.25 (例 16.20 续) 求 $\mathrm{E}(X_3|0,1)$ 的 Bühlmann 估计.

解 从之前的计算中得到

$$\mu(G)=\mathrm{E}(X_j|G)=0.4,\quad \mu(B)=\mathrm{E}(X_j|B)=0.7,$$
$$\pi(G)=0.75,\quad \pi(B)=0.25,$$

因此有

$$\mu=\sum_\theta \mu(\theta)\pi(\theta)=0.4(0.75)+0.7(0.25)=0.475,$$
$$a=\sum_\theta \mu(\theta)^2\pi(\theta)-\mu^2=0.16(0.75)+0.49(0.25)-0.475^2=0.016\ 875.$$

条件方差

$$v(G)=\mathrm{Var}(X_j|G)=0^2(0.7)+1^2(0.2)+2^2(0.1)-0.4^2=0.44,$$
$$v(B)=\mathrm{Var}(X_j|B)=0^2(0.5)+1^2(0.3)+2^2(0.2)-0.7^2=0.61,$$
$$v=\sum_\theta v(\theta)\pi(\theta)=0.44(0.75)+0.61(0.25)=0.482\ 5.$$

由 (16.44) 式可得

$$k=\frac{v}{a}=\frac{0.482\ 5}{0.016\ 875}=28.592\ 6,$$

并且由 (16.43) 式可知

$$Z=\frac{2}{2+28.592\ 6}=0.065\ 4.$$

故下一个索赔额的期望是 0.0654(0.5)+0.9346(0.475)=0.4766. 这就是例 16.20 中给出的贝叶斯保费的最佳线性近似. □

例 16.26 类似例 16.23 的假设 (令 $m_j=1$), $X_j|\Theta$, $j=1,2,\cdots,n$ 独立同分布, 是均值为 Θ 的 Poisson 随机变量, 且 Θ 服从参数为 α,β 的 gamma 分布. 求 Bühlmann 保费.

解 因为

$$\mu(\theta)=\mathrm{E}(X_j|\Theta=\theta)=\theta,\quad v(\theta)=\mathrm{Var}(X_j|\Theta=\theta)=\theta,$$

并且有

$$\mu=\mathrm{E}[\mu(\Theta)]=\mathrm{E}(\theta)=\alpha\beta,\quad v=\mathrm{E}[v(\Theta)]=\mathrm{E}(\Theta)=\alpha\beta,$$

以及

$$a=\mathrm{Var}[\mu(\Theta)]=\mathrm{Var}(\Theta)=\alpha\beta^2.$$

故有

$$k=\frac{v}{a}=\frac{\alpha\beta}{\alpha\beta^2}=\frac{1}{\beta},\quad Z=\frac{n}{n+k}=\frac{n}{n+1/\beta}=\frac{n\beta}{n\beta+1},$$

从而信度保费是

$$Z\bar{X}+(1-Z)\mu=\frac{n\beta}{n\beta+1}\bar{X}+\frac{1}{n\beta+1}\alpha\beta.$$

参见例 16.23 可知, 信度保费就是贝叶斯保费 $\mathrm{E}(X_{n+1}|\boldsymbol{X})$. □

例 16.27 在例 16.22 中求 Bühlmann 保费.

解 对这个模型, 计算可得

$$\begin{aligned}
\mu(\Theta)&=\Theta^{-1},\mu=\mathrm{E}(\Theta^{-1})=\frac{\beta}{\alpha-1},\\
v(\Theta)&=\Theta^{-2},v=\mathrm{E}(\Theta^{-2})=\frac{\beta^2}{(\alpha-1)(\alpha-2)},\\
a&=\mathrm{Var}(\Theta^{-1})=\frac{\beta^2}{(\alpha-1)(\alpha-2)}-\left(\frac{\beta}{\alpha-1}\right)^2=\frac{\beta^2}{(\alpha-1)^2(\alpha-2)},\\
k&=\frac{v}{a}=\alpha-1,\\
Z&=\frac{n}{n+k}=\frac{n}{n+\alpha-1},\\
P_c&=\frac{n}{n+\alpha-1}\bar{X}+\frac{\alpha-1}{n+\alpha-1}\frac{\beta}{\alpha-1},
\end{aligned}$$

计算结果与贝叶斯保费相同.

另一种办法是把所有观察值合并成一个单独的 $S=X_1+\cdots+X_n$. 由题目假定可知 S 的均值是 $n\Theta^{-1}$, 方差是 $n\Theta^{-2}$. 虽然可以证明 S 服从 gamma 分布, 但是在 Bühlmann 近似中不需要用到关于分布的信息, 只用到各阶矩的值. 接下来可以得到

$$\mu=\frac{n\beta}{\alpha-1},\qquad v=\frac{n\beta^2}{(\alpha-1)(\alpha-2)},\qquad a=\frac{n^2\beta^2}{(\alpha-1)^2(\alpha-2)},$$
$$k=\frac{\alpha-1}{n},\qquad Z=\frac{1}{1+k}=\frac{n}{n+\alpha-1}.$$

注意现在样本量变成了 1, 也就是 S 的单个观察值. 由于 $S=n\bar{X}$, 故 Bühlmann 保费是

$$P_c=\frac{n}{n+\alpha-1}n\bar{X}+\frac{\alpha-1}{n+\alpha-1}\frac{n\beta}{\alpha-1},$$

答案是前一个的 n 倍. 因为现在估计的是下一个 S 的值而不是下一个 X 的值. 信度因子 Z 却保持不变, 不论是预测 X_{n+1} 还是下一个 S. □

16.4.5 Bühlmann-Straub 模型

之前提到的 Bühlmann 模型是最简单的信度模型, 它假设投保人过去的索赔数据按年度划分是独立同分布的. 如果在实际应用中允许风险量或样本大小的变动, 那么该假设将不满足.

例如，由于某些特殊原因，某投保人当年的索赔数据只记录了一部分，或者是在保单年当中发生了费率的改变，又或是一个投保群体中投保人数发生了变化，对这些情况应该如何处理呢？

下面考虑 Bühlmann 模型的一般化. 假设 $X_1, \cdots, X_n$ 在给定 Θ 的条件下相互独立，有相同的均值

$$\mu(\theta) = \mathrm{E}(X_j|\Theta = \theta),$$

和条件方差

$$\mathrm{Var}(X_j|\Theta = \theta) = \frac{v(\theta)}{m_j},$$

其中 m_j 是已知常数，是对风险量大小的一个衡量. 实际上，m_j 只需和风险大小成比例即可. 如果把每个 X_j 看成 m_j 个在给定 Θ 的条件下相互独立，且有相同的均值 $\mu(\theta)$ 和方差 $v(\theta)$ 的随机变量的平均，那么该模型的假设可以认为是合理的. 在前述情况中，m_j 可以表示保单在第 j 年中有效的月份数，或是第 j 年该保单的保费收入，或是某群体在第 j 年的投保人数.

和 Bühlmann 模型类似，记

$$\mu = \mathrm{E}[\mu(\Theta)], \quad v = \mathrm{E}[v(\Theta)], \quad a = \mathrm{Var}[\mu(\Theta)].$$

则在无条件的情况下，从 (16.39) 可得 $\mathrm{E}(X_j) = \mu$，从 (16.41) 可得 $\mathrm{Cov}(X_i, X_j) = a$，另外有

$$\begin{aligned}\mathrm{Var}(X_j) &= \mathrm{E}[\mathrm{Var}(X_j|\Theta)] + \mathrm{Var}[\mathrm{E}(X_j|\Theta)] \\ &= \mathrm{E}\left[\frac{v(\Theta)}{m_j}\right] + \mathrm{Var}[\mu(\Theta)] = \frac{v}{m_j} + a.\end{aligned}$$

为了计算信度保费 (16.33)，需要求解正交方程组以确定 $\tilde{\alpha}_0, \tilde{\alpha}_1, \cdots, \tilde{\alpha}_n$. 方便起见，定义

$$m = m_1 + m_2 + \cdots + m_n$$

为总的风险量. 利用 (16.39)，无偏方程 (16.30) 化为

$$\mu = \tilde{\alpha}_0 + \sum_{j=1}^{n} \tilde{\alpha}_j \mu,$$

这等价于

$$\sum_{j=1}^{n} \tilde{\alpha}_j = 1 - \frac{\tilde{\alpha}_0}{\mu}. \tag{16.45}$$

对 $i=1,2,\cdots,n$, (16.32) 可化为

$$a=\sum_{\substack{j=1\\j\neq i}}^{n}\tilde{\alpha}_j a+\tilde{\alpha}_i\left(a+\frac{v}{m_i}\right)=\sum_{j=1}^{n}\tilde{\alpha}_j a+\frac{v\tilde{\alpha}_i}{m_i},$$

也就是

$$\tilde{\alpha}_i=\frac{a}{v}m_i\left(1-\sum_{j=1}^{n}\tilde{\alpha}_j\right)=\frac{a}{v}\frac{\tilde{\alpha}_0}{\mu}m_i,\quad i=1,\cdots,n. \tag{16.46}$$

将其与 (16.45) 相结合, 可得

$$1-\frac{\tilde{\alpha}_0}{\mu}=\sum_{j=1}^{n}\tilde{\alpha}_j=\sum_{i=1}^{n}\tilde{\alpha}_i=\frac{a}{v}\frac{\tilde{\alpha}_0}{\mu}\sum_{i=1}^{n}m_i=\frac{a\tilde{\alpha}_0 m}{\mu v},$$

所以

$$\tilde{\alpha}_0=\frac{\mu}{1+am/v}=\frac{v/a}{m+v/a}\mu.$$

进而得到

$$\tilde{\alpha}_j=\frac{a\tilde{\alpha}_0}{\mu v}\cdot m_j=\frac{m_j}{m+v/a}.$$

于是信度保费 (16.33) 变为

$$\tilde{\alpha}_0+\sum_{j=1}^{n}\tilde{\alpha}_j X_j=Z\bar{X}+(1-Z)\mu, \tag{16.47}$$

其中由 (16.44)$k=v/a$ 知

$$Z=\frac{m}{m+k},\quad \bar{X}=\sum_{j=1}^{n}\frac{m_j}{m}X_j. \tag{16.48}$$

明显看出, 信度保费 (16.47) 也具有 (16.19) 式的形式. 在这里 m 是该投保人的总风险量, 而 Bühlmann-Straub 信度因子 Z 和 m 有关. 另外, $\bar{X}$ 是 X_j 的加权平均, 权重与 m_j 成比例. 从群体的角度看, X_j 是第 j 年 m_j 个成员的平均损失额, m_jX_j 就是该群体在第 j 年的总损失. 从而 $\bar{X}$ 是 n 年中每个成员的总平均损失, 于是对该群体 $n+1$ 年应收取的信度保费是 $m_{n+1}[Z\bar{X}+(1-Z)\mu]$, 其中 m_{n+1} 是下一年的投保人数.

如果事先知道 (16.48) 式就是 X_j 的适当加权, 那么接下来得到信度权重 Z 就比较容易了. 对单个观察值 $\bar{X}$, 它的条件方差是

$$\mathrm{Var}(\bar{X}|\theta)=\sum_{j=1}^{n}\frac{m_j^2}{m^2}\frac{v(\theta)}{m_j}=\frac{v(\theta)}{m}.$$

故条件方差的期望是 v/m. 条件期望的方差仍是 a, 因此 $k = v/(am)$. 现在只有 $\bar{X}$ 的单个观察值, 所以信度因子是

$$Z = \frac{1}{1+v/(am)} = \frac{m}{m+v/a}, \tag{16.49}$$

结果相同. 等式 (16.48) 并不让人感到意外, 因为权重和每个 X_j 的条件方差成简单的反比例关系.

例 16.28 如例 16.23 所假设, 在第 j 年的 m_j 张保单中发生了 N_j 次索赔, $j = 1,2,\cdots,n$. 单独一份保单的索赔次数服从参数为 Θ 的 Poisson 分布, 其中 Θ 服从参数为 α,β 的 gamma 分布, 如果第 $n+1$ 年有 m_{n+1} 份保单, 求索赔次数的 Bühlmann-Straub 估计.

解 设 $X_j = N_j/m_j$, 由于 N_j 服从均值为 $m_j\Theta$ 的 Poisson 分布, 所以有 $\mathrm{E}(X_j|\Theta) = \Theta = \mu(\Theta)$ 和 $\mathrm{Var}(X_j|\Theta) = \Theta/m_j = v(\Theta)/m_j$, 进而可得

$$\begin{aligned} &\mu = \mathrm{E}(\Theta) = \alpha\beta, \quad a = \mathrm{Var}(\Theta) = \alpha\beta^2, \quad v = \mathrm{E}(\Theta) = \alpha\beta, \\ &k = \frac{1}{\beta}, \quad Z = \frac{m}{m+1/\beta} = \frac{m\beta}{m\beta+1}, \end{aligned}$$

对单份保单的估计是

$$P_c = \frac{m\beta}{m\beta+1}\bar{X} + \frac{1}{m\beta+1}\alpha\beta,$$

其中 $\bar{X} = m^{-1}\sum\limits_{j=1}^{n} m_j X_j$. 对第 $n+1$ 年的估计是 $m_{n+1}P_c$, 和例 16.23 答案一致. □

Bühlmann-Straub 模型中作了较强的假设, 这使其在现实中的应用范围受到限制. Hewitt 在 1967 年的一篇文章 [55] 中提到, 大规模风险的表现并非与许多相互独立的小规模风险聚合的表现相一致, 实际上由于独立性它的波动性要更大一些. 下面给出这样一个例子.

例 16.29 设条件期望 $\mathrm{E}(X_j|\Theta) = \mu(\Theta)$, 条件方差 $\mathrm{Var}(X_j|\Theta) = w(\Theta) + v(\Theta)/m_j$. 进一步假定在给定 Θ 的条件下 $X_1,\cdots,X_n$ 相互独立, 证明该模型符合 Hewitt 的观察, 并求信度保费.

解 考虑有相同 Θ 的条件下独立的风险 i 和 j, 分别对应风险量 m_i 和 m_j. 聚合时, 平均损失的方差是

$$\begin{aligned} \mathrm{Var}\left(\frac{m_iX_i+m_jX_j}{m_i+m_j}\middle|\Theta\right) &= \left(\frac{m_i}{m_i+m_j}\right)^2 \mathrm{Var}(X_i|\Theta) + \left(\frac{m_j}{m_i+m_j}\right)^2 \mathrm{Var}(X_j|\Theta) \\ &= \frac{m_i^2+m_j^2}{(m_i+m_j)^2}w(\Theta) + \frac{1}{m_i+m_j}v(\Theta). \end{aligned}$$

而风险量是 m_i+m_j 的单个风险的方差是 $w(\Theta)+v(\Theta)/(m_i+m_j)$, 后者要更大一些.

对信度保费, 计算可得

$$\begin{aligned}\mathrm{E}(X_j) &= \mathrm{E}[\mathrm{E}(X_j|\Theta)] = \mathrm{E}[\mu(\Theta)] = \mu,\\ \mathrm{Var}(X_j) &= \mathrm{E}[\mathrm{Var}(X_j|\Theta)] + \mathrm{Var}[\mathrm{E}(X_j|\Theta)]\\ &= \mathrm{E}\left[w(\Theta) + \frac{v(\Theta)}{m_j}\right] + \mathrm{Var}[\mu(\Theta)] = w + \frac{v}{m_j} + a,\end{aligned}$$

且对 $i \neq j$, $\mathrm{Cov}(X_i, X_j) = a$, 与 (16.41) 一致. 无偏方程仍然是

$$\mu = \tilde{\alpha}_0 + \sum_{j=1}^{n} \tilde{\alpha}_j \mu,$$

故有

$$\sum_{j=1}^{n} \tilde{\alpha}_j = 1 - \frac{\tilde{\alpha}_0}{\mu}.$$

方程 (16.32) 变成

$$\begin{aligned}a &= \sum_{j=1}^{n} \tilde{\alpha}_j a + \tilde{\alpha}_i \left(w + \frac{v}{m_i}\right)\\ &= a\left(1 - \frac{\tilde{\alpha}_0}{\mu}\right) + \tilde{\alpha}_i \left(w + \frac{v}{m_i}\right), \quad i = 1, \cdots, n.\end{aligned}$$

因此

$$\tilde{\alpha}_i = \frac{a\tilde{\alpha}_0/\mu}{w + v/m_i}.$$

两边求和得到

$$\frac{a\tilde{\alpha}_0}{\mu} \sum_{j=1}^{n} \frac{m_j}{v + wm_j} = \sum_{j=1}^{n} \tilde{\alpha}_j = 1 - \frac{\tilde{\alpha}_0}{\mu},$$

所以

$$\tilde{\alpha}_0 = \frac{1}{(a/\mu)\sum_{j=1}^{n} \frac{m_j}{v+wm_j} + \frac{1}{\mu}} = \frac{\mu}{1 + am^*},$$

其中

$$m^* = \sum_{j=1}^{n} \frac{m_j}{v + wm_j}.$$

进而有

$$\tilde{\alpha}_j = \frac{am_j}{v + wm_j} \frac{1}{1 + am^*}.$$

信度保费是

$$\frac{\mu}{1 + am^*} + \frac{a}{1 + am^*} \sum_{j=1}^{n} \frac{m_j X_j}{v + wm_j}.$$

定义观察值的加权平均为

$$\bar{X} = \frac{\sum_{j=1}^{n} \frac{m_j}{v+wm_j} X_j}{\sum_{j=1}^{n} \frac{m_j}{v+wm_j}} = \frac{1}{m^*} \sum_{j=1}^{n} \frac{m_j}{v+wm_j} X_j.$$

再令

$$Z = \frac{am^*}{1+am^*},$$

则信度保费是

$$Z\bar{X} + (1-Z)\mu.$$

如果风险量 $m_j(j=1,2,\cdots,n)$ 趋于无穷, 则信度因子

$$Z \to \frac{an/w}{1+an/w} < 1.$$

而 Bühlmann-Straub 模型的极限是 1. 也就是说不管风险量多大, 信度因子是有上限的. 在习题 (16.26) 中会对这个结果进行一些推广. □

另一种推广是让 $\mu(\Theta)$ 的方差与风险量有关. 如果相信某给定风险引起索赔的倾向偏离均值的程度与风险自身的大小有关, 这个假设可以认为是合理的. 比如说, 大的风险保单会更加谨慎地签单, 使得风险不但要在签单时满足签单要求, 索赔时还要满足索赔要求, 这样发生偏离均值的极端情形的可能性会更小一些.

例 16.30　(例 16.29 续) 除了例 16.29 的条件外, 设 $\text{Var}[\mu(\Theta)] = a + b/m$, 其中 $m = \sum_{j=1}^{n} m_j$ 是群体总风险量. 试推导信度保费的公式.

解　现在已有

$$\begin{aligned}
\text{E}(X_j) &= \text{E}[\text{E}(X_j|\Theta)] = \text{E}[\mu(\Theta)] = \mu, \\
\text{Var}(X_j) &= \text{E}[\text{Var}(X_j|\Theta)] + \text{Var}[\text{E}(X_j|\Theta)] \\
&= \text{E}\left[w(\Theta) + \frac{v(\Theta)}{m_j}\right] + \text{Var}[\mu(\Theta)] \\
&= w + \frac{v}{m_j} + a + \frac{b}{m},
\end{aligned}$$

且对 $i \neq j$,

$$\begin{aligned}
\text{Cov}(X_i, X_j) &= \text{E}[\text{E}(X_i X_j|\Theta)] - \mu^2 \\
&= \text{E}[\mu(\Theta)^2] - \mu^2 = a + \frac{b}{m}.
\end{aligned}$$

可以看到所有例 16.29 中的计算均适用, 只需把 a 换成 $a + b/m$ 即可. 信度因子是

$$Z = \frac{(a+b/m)m^*}{1+(a+b/m)m^*},$$

信度保费是

$$Z\bar{X}+(1-Z)\mu,$$

其中 $\bar{X}$ 和 m^* 如例 16.29 中定义. 这个特定的信度公式曾被用于工伤保险的费率计算. [45] 中给出了一个详细的例子. □

16.4.6 精确信度

例 16.26 至例 16.28 中信度保费和贝叶斯保费相等. 从 (16.34) 式知道, 在均方误差损失的意义下信度保费是贝叶斯保费的最佳线性近似. 在这些例子中因为两种保费相等, 所以近似是精确的, 用术语**精确信度**来表示信度保费等于贝叶斯保费的情形.

实际上在不计算出信度保费的情形下, 也有可能知道信度保费是否精确. 如果贝叶斯保费是 $X_1,\cdots,X_n$ 的线性函数,

$$\mathrm{E}(X_{n+1}|\boldsymbol{X})=a_0+\sum_{j=1}^{n}a_jX_j,$$

则在 (16.34) 式中 Q_1 在 $\tilde{\alpha}_j=\alpha_j,\ j=0,1,\cdots,n$ 的情形下取到最小值 0, 因此信度保费就是 $\tilde{\alpha}_0+\sum\limits_{j=1}^{n}\tilde{\alpha}_jX_j=\alpha_0+\sum\limits_{j=1}^{n}\alpha_jX_j=\mathrm{E}(X_{n+1}|\boldsymbol{X})$, 它是精确的.

这种情形在与线性指数分布族及其共轭先验分布相联系时经常发生 (12.4.3 节). 假设 $X_j|\Theta=\theta$(在 $\Theta=\theta$ 条件下) 相互独立, 且对 $\mathrm{j}=1,\cdots,n+1$ 分别有概率密度

$$f_{X_j|\Theta}(x_j|\theta)=\frac{p(x_j)\mathrm{e}^{-\theta x_j}}{q(\theta)},$$

Θ 有密度函数

$$\pi(\theta)=\frac{[q(\theta)]^{-k}e^{-\mu k\theta}}{c(\mu,k)},\quad \theta_0<\theta<\theta_1, \tag{16.50}$$

其中 $-\infty\leqslant\theta_0<\theta_1\leqslant\infty$. 同时假定 $\pi(\theta_0)=\pi(\theta_1)=0$. 现在 μ,k 看上去仅仅是 $\pi(\theta)$ 的参数, 下面将说明符号的选择并不是巧合.

在 12.4.3 节中证明了

$$\mu(\theta)=\mathrm{E}(X_j|\Theta=\theta)=-\frac{q'(\theta)}{q(\theta)}.$$

希望计算 $E[\mu(\Theta)]$. 由 (16.50) 式得

$$\ln\pi(\theta)=-k\ln q(\theta)-\mu k\theta-\ln c(\mu,k),$$

对 θ 求导, 有

$$\frac{\pi'(\theta)}{\pi(\theta)}=-\frac{kq'(\theta)}{q(\theta)}-\mu k.$$

也就是

$$\pi'(\theta) = k[\mu(\theta) - \mu]\pi(\theta), \tag{16.51}$$

再从 θ_0 到 θ_1 积分, 得到

$$\pi(\theta_1) - \pi(\theta_0) = k\int_{\theta_0}^{\theta_1} \mu(\theta)\pi(\theta)\mathrm{d}\theta - k\mu\int_{\theta_0}^{\theta^1} \pi(\theta)\mathrm{d}\theta.$$

这等价于 $0=k\mathrm{E}[\mu(\Theta)] - k\mu$, 即

$$\mathrm{E}[\mu(\Theta)] = \mu. \tag{16.52}$$

现在来计算后验分布 $\pi_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$. 它与

$$\left[\prod_{j=1}^{n} f_{X_j|\Theta}(x_j|\theta)\right]\pi(\theta),$$

成比例, 也就是与

$$\begin{aligned}\left[\prod_{j=1}^{n} \frac{\mathrm{e}^{-\theta x_j}}{q(\theta)}\right][q(\theta)]^{-k}\mathrm{e}^{-\mu k\theta} &= [q(\theta)]^{-(n+k)}\mathrm{e}^{-\theta(\mu k+n\bar{x})}\\ &= [q(\theta)]^{-k_*}\mathrm{e}^{-\mu_* k_*\theta},\end{aligned} \tag{16.53}$$

成比例, 其中

$$k_* = n+k, \quad \mu_* = \frac{\mu k + n\bar{x}}{k+n} = \frac{n}{n+k}\bar{x} + \frac{k}{n+k}\mu.$$

注意到 (16.53) 式与 (16.50) 式的密度形式 (把 μ, k 分别换成 μ_*, k_*) 成比例关系, 因此有

$$\pi_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x}) = \frac{[q(\theta)]^{-k_*}\mathrm{e}^{-\mu_* k_*\theta}}{c(\mu_*, k_*)}, \quad \theta_0 < \theta < \theta_1.$$

利用 (16.28) 式和与推导 (16.52) 式相同的方法, 得到贝叶斯保费是

$$\mathrm{E}(X_{n+1|\boldsymbol{X}}) = \int_{\theta_0}^{\theta_1} \mu(\theta)\pi_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})\mathrm{d}\theta = \mu_* = Z\bar{x} + (1-Z)\mu,$$

其中 $Z = n/(n+k)$. 这就是 (16.19) 的形式, 并且因为它是诸 x_j 的线性函数, 所以必然满足精确信度, 即信度保费是

$$\tilde{\alpha}_0 + \sum_{j=1}^{n} \tilde{\alpha}_j X_j = Z\bar{X} + (1-Z)\mu = \mathrm{E}(X_{n+1}|\boldsymbol{x}).$$

同时由于 $X_j|\Theta(j=1,\cdots,n)$ 是同分布的, 适用 Bühlmann 模型, 故 (16.42) 式也适用. 那就是说 k 必然满足 (16.44) 式. 为了直接说明这一点, 回忆在 12.4.3 节中有

$$v(\theta)=\mathrm{Var}(X_j|\Theta=\theta)=-\mu'(\theta).$$

对 (16.51) 式求导可得

$$\begin{aligned}\pi''(\theta)&=k\mu'(\theta)\pi(\theta)+k^2[\mu(\theta)-\mu]^2\pi(\theta)\\&=-kv(\theta)\pi(\theta)+k^2[\mu(\theta)-\mu]^2\pi(\theta).\end{aligned}$$

对 θ 从 θ_0 到 θ_1 积分, 有

$$\pi'(\theta_1)-\pi'(\theta_0)=-k\mathrm{E}[v(\Theta)]+k^2\mathrm{E}\{[\mu(\Theta)-\mu]^2\}=-kv+k^2a.$$

这是因为 $\mu(\Theta)$ 的期望是 μ, 且 $\mathrm{E}\{[\mu(\Theta)-\mu]^2\}=\mathrm{Var}[\mu(\Theta)]=a$. 如果 $\pi'(\theta_1)=\pi'(\theta_0)=0$, 则有 $k=v/a$, 满足 (16.44) 式.

16.4.7 线性保费, 贝叶斯保费和无信度之间的比较

在 16.4.3 节中证明了信度保费是在最小均方误差的意义下 X_{n+1} 的最佳线性估计. 在习题 16.59 中将要求读者证明, 在没有其他条件的限制下, 贝叶斯保费在所有 X_{n+1} 的估计量中是均方误差最小的. 在 16.4.3 节中还证明了在均方误差意义下信度保费是贝叶斯保费的最佳线性近似, 并且在若干个例子中, 两者完全相等. 这就产生了两个问题. 一个是, 在使用信度保费时, 一般说来会比贝叶斯保费产生更大的误差, 这些增加的误差值得我们担心吗? 另一个是, 是否要在意采用信度方法所带来的困惑. 要想准确地回答以上问题, 必须就事论事地考察数据本身服从的概率分布. 下面希望通过两个例子, 给读者带来一些感性的认识.

下面将利用一个已讨论过的情形, 先从第二个问题开始. 因为要进行多次估计, 所以信度理论才有用武之地. 之所以在某些情况下使用这种有偏估计, 是因为偏差能够随着估计次数的增加而逐渐抵消, 也就是说, 波动性或平方误差将会减少. 下面的例子显示了信度理论的作用.

例 16.31 假设在 50 种不同的条件下各自随机地得到一组含 10 个观察值的样本, 它们均来自均值未知的 Poisson 分布, 且每组样本的 Poisson 分布均值可以互不相同. 设这些均值的真值是 $\theta_1,\cdots,\theta_{50}$, 并进一步假定这些 Poisson 参数都是选自参数 $\alpha=50,\beta=0.1$ 的 gamma 分布. 比较最大似然估计 $\bar{X}_j, j=1,\cdots,50$ 和信度估计 $C_j=(\bar{X}_j+5)/2$(Bühlmann 信度估计).

解 首先通过计算各自的均方误差来分析这两个估计量. 样本均值的总平方误差是

$$S_1 = \sum_{j=1}^{50} (\bar{X}_j - \Theta_j)^2,$$

其中 $\boldsymbol{\Theta} = (\Theta_1, \cdots, \Theta_{50})$. 从而均方误差是

$$\mathrm{E}(S_1) = \mathrm{E}[\mathrm{E}(S_1|\boldsymbol{\Theta})] = \mathrm{E}\left[\sum_{j=1}^{50} \mathrm{Var}(\bar{X}_j|\Theta_j)\right] = \mathrm{E}\left(\sum_{j=1}^{50} \frac{\Theta_j}{10}\right) = 25.$$

信度估计的平方误差是

$$S_2 = \sum_{j=1}^{50} (0.5\bar{X}_j + 2.5 - \Theta_j)^2.$$

均方误差是

$$\begin{aligned}
\mathrm{E}(S_2) &= \mathrm{E}[\mathrm{E}(S_2|\boldsymbol{\Theta})] \\
&= \mathrm{E}\left[\sum_{j=1}^{50} E(0.25\bar{X}_j^2 + 6.25 + \Theta_j^2 + 2.5\bar{X}_j - 5\Theta_j - \bar{X}_j\Theta_j|\Theta_j)\right] \\
&= \mathrm{E}\left\{\sum_{j=1}^{50} \left[0.25\left(\frac{\Theta_j}{10} + \Theta_j^2\right) + 6.25 + \Theta_j^2 + 2.5\Theta_j - 5\Theta_j - \Theta_j^2\right]\right\} \\
&= \sum_{j=1}^{50} [0.25(0.5 + 25.5) + 6.25 + 25.5 + 2.5(5) - 5(5) - 25.5] \\
&= 12.5.
\end{aligned}$$

这个例子有误导的嫌疑, 我们事先知道 Bühlmann 估计在所有线性估计中拥有最小的均方误差, 却又人为地把平方误差作为比较的准则. 有意思的是结果有了明显改善. 这就表明, 即使在使用信度公式中 Z 和 μ 的取值不够准确, 信度保费仍然有可能带来均方误差的减少.

为了直观地理解这种改善从何而来, 考虑具体的 50 个 θ_j. 表 16-3 是某个 gamma 分布的样本, 按从小到大排序. 第二列是关于样本均值 $(\theta_j/10)$ 的均方误差, 最后三列分别是信度估计 ($Z = 0.5$ 和 $\mu = 0.5$) 的期望偏差, 方差和均方误差. 由于样本均值是无偏的, 所以期望偏差是 0, 同时方差也就是均方误差, 所以并没有给出这两项数据.

对信度估计有

期望偏差 $= \mathrm{E}(0.5\bar{X}_j + 2.5 - \theta_j) = 2.5 - 0.5\theta_j$,

方差 $= \mathrm{Var}(0.5\bar{X}_j + 2.5) = 0.25\theta_j/10 = 0.025\theta_j$,

均方误差 $=$ 期望偏差的平方 $+$ 方差 $= 0.25\theta_j^2 - 2.475\theta_j + 6.25$.

表 16-3 样本均值和信度估计的比较

Θ	$\bar{X}$	$0.5\bar{X}+2.5$			θ	$\bar{X}$	$0.5\bar{X}+2.5$		
	均方误差	偏差	方差	均方误差		均方误差	偏差	方差	均方误差
3.510	0.351	0.745	0.088	0.643	4.875	0.488	0.062	0.122	0.126
3.637	0.364	0.681	0.091	0.555	4.894	0.489	0.053	0.122	0.125
3.742	0.374	0.629	0.094	0.489	4.900	0.490	0.050	0.123	0.125
3.764	0.376	0.618	0.094	0.476	4.943	0.494	0.028	0.124	0.124
3.793	0.379	0.604	0.095	0.459	4.977	0.498	0.012	0.124	0.125
4.000	0.400	0.500	0.100	0.350	5.002	0.500	−0.001	0.125	0.125
4.151	0.415	0.424	0.104	0.284	5.013	0.501	−0.006	0.125	0.125
4.153	0.415	0.424	0.104	0.283	5.108	0.511	−0.054	0.128	0.131
4.291	0.429	0.354	0.107	0.233	5.172	0.517	−0.086	0.129	0.137
4.405	0.440	0.298	0.110	0.199	5.198	0.520	−0.099	0.130	0.140
4.410	0.441	0.295	0.110	0.197	5.231	0.523	−0.116	0.131	0.144
4.413	0.441	0.293	0.110	0.196	5.239	0.524	−0.120	0.131	0.145
4.430	0.443	0.285	0.111	0.192	5.263	0.526	−0.132	0.132	0.149
4.438	0.444	0.281	0.111	0.190	5.300	0.530	−0.150	0.132	0.155
4.471	0.447	0.264	0.112	0.182	5.338	0.534	−0.169	0.133	0.162
4.491	0.449	0.254	0.112	0.177	5.400	0.540	−0.200	0.135	0.175
4.495	0.449	0.253	0.112	0.176	5.407	0.541	−0.203	0.135	0.176
4.505	0.451	0.247	0.113	0.174	5.431	0.543	−0.215	0.136	0.182
4.547	0.455	0.227	0.114	0.165	5.459	0.546	−0.229	0.136	0.189
4.606	0.461	0.197	0.115	0.154	5.510	0.551	−0.255	0.138	0.203
4.654	0.465	0.173	0.116	0.146	5.538	0.554	−0.269	0.138	0.211
4.758	0.476	0.121	0.119	0.134	5.646	0.565	−0.323	0.141	0.246
4.763	0.476	0.118	0.119	0.133	5.837	0.584	−0.419	0.146	0.321
4.766	0.477	0.117	0.119	0.133	5.937	0.594	−0.468	0.148	0.368
4.796	0.480	0.102	0.120	0.130	6.263	0.626	−0.631	0.157	0.555
					均值	0.482	0.091	0.120	0.222

正如预计的那样, 对信度估计而言, 通过允许在单个估计中出现偏差而让均方误差得以显著减少. 进一步地, 信度估计还是最接近先验分布的均值 (5) 的估计量. □

现在已经看到了信度理论的应用价值. 下一个任务是比较线性信度估计和贝叶斯估计. 在大部分例子中是很难比较的, 因为贝叶斯估计只能通过近似积分计算得到. 一种替代办法是通过模拟来计算均方误差, 在 Foundations of Actuarial Science [24], P.467 中有相关的阐述. 在下面的例子中将进行相同的阐述, 不过采用某种近似手段以避免对近似的积分计算. 值得一提的是线性信度保费只需估算前两阶矩, 而贝叶斯方法要求知道具体的分布形式. 非参数的特征使得线性方法能够处理多种复杂情形, 足以弥补不够精确的缺点.

例 16.32 一组观察值的样本量是 25, 来自参数 $\alpha = 4$, 标度参数 Θ 未知的逆 gamma 分布. Θ 的先验分布是均值为 50, 方差为 5 000 的 gamma 分布, 试比较线性估计和贝叶斯估计.

解 对 Bühlmann 线性信度估计有

$$\mu = \mathrm{E}[\mu(\Theta)] = \mathrm{E}\left(\frac{\Theta}{3}\right) = \frac{50}{3},$$

$$a = \mathrm{Var}[\mu(\Theta)] = \mathrm{Var}\left(\frac{\Theta}{3}\right) = \frac{5\,000}{9},$$

$$v = \mathrm{E}[v(\Theta)] = \mathrm{E}\left(\frac{\Theta^2}{18}\right) = \frac{5\,000 + 50^2}{18} = \frac{7\,500}{18},$$

因此有

$$Z = \frac{25}{25 + \frac{7\,500/18}{5\,000/9}} = \frac{100}{103},$$

信度估计值是 $\hat{\mu}_{\text{cred}} = (100\bar{X} + 50)/103$.

若用贝叶斯估计, 后验密度为

$$\pi_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x}) \propto \mathrm{e}^{-\theta\sum_{j=1}^{25} x_j^{-1}}\theta^{100}\theta^{-0.5}\mathrm{e}^{-\theta/100} \propto \theta^{99.5}\mathrm{e}^{-\theta\left(0.01+\sum_{j=1}^{25} x_j^{-1}\right)},$$

它是参数为 100.5 和 $\left(0.01 + \sum\limits_{j=1}^{25} x_j^{-1}\right)^{-1}$ 的 gamma 分布的密度函数. 于是后验均值是

$$\hat{\theta}_{\text{Bayes}} = \frac{100.5}{0.01 + \sum_{j=1}^{25} x_j^{-1}},$$

因此

$$\hat{\mu}_{\text{Bayes}} = \frac{33.5}{0.01 + \sum_{j=1}^{25} x_j^{-1}},$$

显然它是非线性估计量.

至于不同估计的精确程度, 可以附带考虑样本均值估计量. 在给定 θ 的前提下样本均值是无偏的, 且方差和均方误差都是 $\theta^2/(18 \times 25) = \theta^2/450$; 信度估计的期望偏差是

$$\begin{aligned}\mathrm{bias}_\theta(\hat{\mu}_{\text{cred}}) &= \mathrm{E}\left(\frac{100\bar{X}}{103} + \frac{50}{103} - \frac{\theta}{3}\right)\\ &= \frac{100\theta}{309} + \frac{50}{103} - \frac{\theta}{3} = \frac{50}{103} - \frac{\theta}{103},\end{aligned}$$

方差是

$$\mathrm{Var}_\theta(\hat{\mu}_{\text{cred}}) = \frac{(100/103)^2\theta^2}{450},$$

均方误差是

$$\mathrm{MSE}_\theta(\hat{\mu}_{\text{cred}}) = \frac{1}{103^2}\left(2\ 500 - 100\theta + \frac{10\ 450\theta^2}{450}\right).$$

在计算贝叶斯估计时, 若给定 θ, 则 $1/X$ 服从参数为 4 和 $1/\theta$ 的 gamma 分布, 因此 $\sum_{j=1}^{25} x_j^{-1}$ 服从参数为 100 和 $1/\theta$ 的 gamma 分布. 注意到 $\hat{\mu}_{\text{Bayes}}$ 的分母中的 0.01 占整个分母的比例常常会很小, 因此不妨略去 0.01, 即近似认为 $\hat{\mu}_{\text{Bayes}}$ 服从参数为 100 和 33.5θ 的逆 gamma 分布. 这时有

$$\mathrm{Bias}_\theta(\hat{\mu}_{\text{Bayes}}) = \frac{33.5\theta}{99} - \frac{\theta}{3} = \frac{0.5\theta}{99},$$

$$\mathrm{Var}_\theta(\hat{\mu}_{\text{Bayes}}) = \frac{33.5^2\theta^2}{99^2(98)},$$

$$\mathrm{MSE}_\theta(\hat{\mu}_{\text{Bayes}}) = \frac{33.5^2 + 49/2}{99^2(98)}\theta^2 = 0.001\ 193\ 91\theta^2.$$

三个估计量: 样本均值, 信度估计和贝叶斯估计, 它们的均方误差中 θ^2 的系数分别是 0.002 22,0.002 19,0.001 19. 因此对较大的 θ 而言信度估计的均方误差并没有明显小于样本均值, 但是贝叶斯估计让均方误差减少了将近一半. 在表 16-4 中给出了对 gamma 先验分布不同百分位点的各种量的值. □

表 16-4 样本均值, 信度估计和贝叶斯估计的比较

分位点	θ	$\bar{X}$	$\hat{\mu}_{\text{cred}}$		$\hat{\mu}_{\text{Bayes}}$	
		均方误差	偏差	均方误差	偏差	均方误差
1	0.008	0.000	0.485	0.236	0.000	0.000
5	0.197	0.000	0.484	0.234	0.001	0.000
10	0.790	0.001	0.478	0.230	0.004	0.001
25	5.077	0.057	0.436	0.244	0.026	0.031
50	22.747	1.150	0.265	1.154	0.115	0.618
75	66.165	9.729	−0.157	9.195	0.334	5.227
90	135.277	40.667	−0.828	39.018	0.683	21.849
95	192.072	81.982	−1.379	79.178	0.970	44.046
99	331.746	244.568	−2.735	238.011	1.675	131.397

和贝叶斯估计相比, 信度估计表现较差的原因是题中的 gamma 分布有较厚的尾. 一种削减尾部的办法是对数据取对数. 这个想法在 [24] 中提出, 并对上述例子进行了计算. 做法是先对数据取对数, 然后用线性信度去估计取对数后数据的均值, 最后把估计值取幂. 由于这种做法必然会带来偏差①, 因此需要略微放大估计值. 下面的例子就是用这种方法, 许多细节留到了习题 16.57.

① 由 Jensen 不等式, $\mathrm{E}[\ln X] < \ln \mathrm{E}(X)$, 因此取对数的操作会低估了实际值.

例 16.33　(例 16.32 续) 求解对数信度估计量, 并计算其偏差和均方误差.

解　定义 $W_j = \ln X_j$, 用取对数后的数据进行计算, 有

$$\mu(\Theta) = \mathrm{E}(W|\Theta) = \int_0^\infty (\ln x)\Theta^4 x^{-5} \mathrm{e}^{-\Theta/x} \frac{1}{6} \mathrm{d}x$$
$$= \int_0^\infty (\ln\Theta - \ln y) y^3 \mathrm{e}^{-y} \frac{1}{6} \mathrm{d}y = \ln\Theta - \Psi(4),$$

其中对第二个等号后的积分作了变量替换 $y = \Theta/x$. 最后一个等号成立是因为 $y^3\mathrm{e}^{-y}/6$ 是 gamma 分布的密度函数, 所以积分是 1, 同时第二项是**digamma 函数**(见习题 16.57). 查 [3] 中的表可得 $\Psi(4) = 1.25612$. 下一个要求的量是

$$\begin{aligned} v(\Theta) &= \mathrm{E}(W^2|\theta) - \mu(\Theta)^2 \\ &= \int_0^\infty (\ln x)^2 \Theta^4 x^{-5} \mathrm{e}^{-\Theta/x} \frac{1}{6} \mathrm{d}x - [\ln\Theta - \Psi(4)]^2 \\ &= \int_0^\infty (\ln\Theta - \ln y)^2 y^3 \mathrm{e}^{-y} \frac{1}{6} \mathrm{d}x - [\ln\Theta - \Psi(4)]^2 = \Psi'(4), \end{aligned}$$

其中 $\psi'(4)$=0.283 823 是**trigamma 函数** (见习题 16.57). 故有

$$\begin{aligned} \mu &= \mathrm{E}[\ln\Theta - \Psi(4)] \\ &= \int_0^\infty (\ln\theta)\theta^{-0.5} \mathrm{e}^{-\theta/100} 100^{-0.5} \frac{1}{\Gamma(0.5)} \mathrm{d}\theta - \Psi(4) \\ &= \int_0^\infty (\ln 100 + \ln\lambda)\lambda^{-0.5} \mathrm{e}^{-\lambda} \frac{1}{\Gamma(0.5)} \mathrm{d}\lambda - \Psi(4) \\ &= \ln 100 + \Psi(0.5) - \Psi(4) = 1.385\ 54. \end{aligned}$$

且

$$v = \mathrm{E}[\Psi'(4)] = \Psi'(4) = 0.283\ 823,$$
$$a = \mathrm{Var}[\ln\Theta - \Psi(4)] = \Psi'(0.5) = 4.934\ 802,$$
$$Z = \frac{25}{25 + \frac{0.283\ 823}{4.934\ 802}} = 0.997\ 705.$$

对数信度估计是

$$\hat{\mu}_{\text{log-cred}} = c\exp(0.997\ 705\bar{W} + 0.003\ 180\ 24).$$

其中 c 由下式确定

$$\begin{aligned} \mathrm{E}(X) = \frac{50}{3} &= c\mathrm{E}[\exp(0.997\ 705\bar{W} + 0.003\ 180\ 24)] \\ &= c\mathrm{e}^{0.003\ 180\ 24} \mathrm{E}\left[\exp\left(\frac{0.997\ 705}{25} \sum_{j=1}^{25} \ln X_j\right)\right] \end{aligned}$$

$$= ce^{0.003\ 180\ 24}\mathrm{E}\left[\mathrm{E}\left(\prod_{j=1}^{25} X_j^{0.997\ 705/25}|\Theta\right)\right].$$

给定 Θ 时诸 X_j 是相互独立的, 所以乘积的期望就是期望的乘积. 从附录 A 关于逆 gamma 分布的第 k 阶矩的公式可知

$$\begin{aligned}\frac{50}{3} &= ce^{0.003\ 180\ 24}\mathrm{E}\left\{\left[\frac{1}{6}\Theta^{0.997\ 705/25}\Gamma\left(4-\frac{0.997\ 705}{25}\right)\right]^{25}\right\}\\ &= ce^{0.003\ 180\ 24}\left[\frac{1}{6}\Gamma\left(4-\frac{0.997\ 705}{25}\right)\right]^{25}\frac{100^{0.997\ 705}\Gamma(0.5+0.997\ 705)}{\Gamma(0.5)},\end{aligned}$$

进而得到 $c = 1.169\ 318$ 以及

$$\hat{\mu}_{\text{log-cred}} = 1.173\ 043(2.712\ 051)^{\bar{W}}.$$

对任意给定的 θ 为了求偏差和均方误差, 需要计算

$$\begin{aligned}\mathrm{E}(\hat{\mu}_{\text{log-cred}}|\Theta=\theta) &= 1.173\ 043\mathrm{E}(e^{\bar{W}\ln 2.712\ 051}|\Theta=\theta)\\ &= 1.173\ 043\mathrm{E}\left[\prod_{j=1}^{25} X_j^{(\ln 2.712\ 051)/25}|\Theta=\theta\right]\\ &= 1.173\ 043\left[\frac{1}{6}\theta^{(\ln 2.712\ 051)/25}\Gamma\left(4-\frac{\ln 2.712\ 051}{25}\right)\right]^{25},\\ \mathrm{E}(\hat{\mu}^2_{\text{log-cred}}|\Theta=\theta) &= 1.173\ 043^2\mathrm{E}(e^{2\bar{W}\ln 2.712\ 051}|\Theta=\theta)\\ &= 1.173\ 043^2\left[\frac{1}{6}\theta^{(2\ln 2.712\ 051)/25}\Gamma\left(4-\frac{2\ln 2.712\ 051}{25}\right)\right]^{25}.\end{aligned}$$

评价估计好坏的量是

$$\begin{aligned}\mathrm{Bias}_\theta(\hat{\mu}_{\text{log-cred}}) &= \mathrm{E}(\hat{\mu}_{\text{log-cred}}|\Theta=\theta)-\frac{1}{3}\theta,\\ \mathrm{MSE}_\theta(\hat{\mu}_{\text{log-cred}}) &= \mathrm{E}(\hat{\mu}^2_{\text{log-cred}}|\Theta=\theta)-[\mathrm{E}(\hat{\mu}_{\text{log-cred}}|\Theta=\theta)]^2+[\mathrm{bias}_\theta(\hat{\mu}_{\text{log-cred}})]^2.\end{aligned}$$

表 16-5 中给出了不同的 θ 所对应的各种量的值. 和表 16-4 比较发现, 对数信度估计几乎与贝叶斯估计一样好. □

在实际中, 对数信度和普通的信度估计一样易于操作, 而且都将用到 16.5 节中介绍的计算方法之一. 在对数信度中, 用数据的对数值代替观察值, 然后把得到的估计量取幂. 估计量的有偏性通过如下方法修正: 把估计量乘上一个常数, 该常数使得估计量的样本均值和原始数据的样本均值相等.

表 16-5 对数信度估计量的偏差和均方误差

分位点	θ	偏差	均方误差
1	0.008	0.000	0.000
5	0.197	0.001	0.000
10	0.790	0.003	0.001
25	5.077	0.012	0.034
50	22.747	0.026	0.666
75	66.165	0.023	5.604
90	135.277	−0.028	23.346
95	192.072	−0.091	46.995
99	331.746	−0.295	139.908

16.4.8 备注

关于限制波动信度理论有两个主要的批评意见, 本节讨论了其中的一个. 通过使用条件期望的方差, 得到了一种方法, 把要研究的群体的均值 $\mu(\theta)$ 和手册 (集体) 保费 μ 联系起来. 当中的推导是正确, 客观的, 有具体的模型背景. 另外也发现, 对估计加上额外的线性限制条件后, 结果并没有想象中那么差, 还不时能得到精确贝叶斯信度的结果. 随后我们花了大量工夫来对原始模型进行一般化, 这为计算信度保费打下了一个良好的基础. 然而, 有一个困难并没有解决: 如何从数值上估计 Bühlmann 公式中的量 a 和 v, 或者是如何把贝叶斯公式中的先验分布具体化? 这些问题将在本章的最后一节中进行论述.

Norberg[100] 给出了关于信度理论发展的一个回顾, 当中有关于限制波动信度理论和最精确信度理论的描述. 从 Bühlmann 的经典论文 [18] 之后, 产生了大量关于信度理论的精算文献, 其他一些基本的介绍性文献有 Herzog[52] 和 Waters[135]. 另外一些进一步的处理可见 Goovaerts and Hoogstad[46] 和 Sundt[127]. Bühlmann-Straub 模型的一个重要的推广是 Hachemeister[48] 的回归模型, 在这里不作讨论. 还可以参见 Klugman[76]. 关于精确信度的论述来自 Jewell[66], 也可见 Ericson[34]. 在某一期的 Insurance: Abstracts and Reviews(Sundt[126]) 还列出了一张信度理论文献的清单.

习题

16.22 考虑如下的骰子-转盘模型. 1 个骰子 1 个面有标记, 5 个面没有标记; 另一个骰子 4 个面有标记, 2 个面没有标记. 有 3 个转盘, 每个都被平均分成 5 个区域, 每个区域上标有数字 3 或 8. 第 1 个转盘有 1 个区是 3, 4 个区是 8; 第 2 个有 2 个 3, 3 个 8; 第 3 个有 4 个 3, 1 个 8. 先随机选定 1 个骰子和 1 个转盘. 如果掷骰子得到没有标记的面, 则没有损失发生, 但如果掷到有标记的面, 则旋转选出的转盘以确定损失额.

(a) 对每个可能的骰子-转盘组合求 $\pi(\theta)$.

(b) 对每种骰子–转盘组合, 计算损失额的条件分布 $f_{X|\Theta}(x|\theta)$.

(c) 对每个 θ, 求对应的条件期望 $\mu(\theta)$ 和条件方差 $v(\theta)$.

(d) 计算首次操作的损失额 X_1 是 3 的边缘概率.

(e) 用贝叶斯公式计算后验概率 $\pi_{\Theta|\mathbf{X}_1}(\theta|3)$.

(f) 已知第一次操作得到 $X_1 = 3$, 利用 (16.25) 求第二次操作损失额 X_2 的条件分布 $f_{X_2|X_1}(x_2|3)$.

(g) 利用 (16.28) 式计算贝叶斯保费 $\mathrm{E}(X_2|X_1 = 3)$.

(h) 分别对 $x_2 = 0, 3, 8$ 求 $X_2 = x_2$ 和 $X_1 = 3$ 的联合概率.

(i) 直接用 (16.23) 式计算条件分布 $f_{X_2|X_1}(x_2|3)$, 并与 (f) 的结果相比较.

(j) 直接用 (16.27) 式求贝叶斯保费, 并和 (g) 的结果相比较.

(k) 确定结构参数 μ, v, a.

(l) 作为贝叶斯保费 $\mathrm{E}(X_2|X_1 = 3)$ 的近似, 计算 Bühlmann 信度因子和 Bühlmann 信度保费.

16.23 3 个罐子中分别有编号是 0, 1, 2 的许多球, 每个编号在不同的罐子中所占比例如表 16-6 所示. 先随机选取 1 个罐子, 再从这个罐子中有放回地取出 2 个球. 已知这 2 个球编号总和是 2. 然后再次从这个罐子中有放回地取出 2 个球, 要估计这 2 个球编号之和.

(a) 确定 $\pi(\theta)$.

(b) 对每个罐子分别计算 2 球编号之和的条件分布 $f_{X|\Theta}(x|\theta)$.

(c) 对每个 θ, 求对应的条件期望 $\mu(\theta)$ 和条件方差 $v(\theta)$.

(d) 计算第 1 次抽取中 2 个球编号之和 X_1 是 2 的边缘概率.

(e) 用贝叶斯公式计算后验概率 $\pi_{\Theta|\mathbf{X}_1}(\theta|2)$.

(f) 已知第 1 次取出的 2 球编号之和 $X_1 = 2$, 利用 (16.25) 求第 2 次取出的 2 球编号之和 X_2 的条件分布 $f_{X_2|X_1}(x_2|2)$.

(g) 利用 (16.28) 式计算贝叶斯保费 $\mathrm{E}(X_2|X_1 = 2)$.

(h) 分别对 $x_2 = 0, 1, 2, 3, 4$ 求 $X_2 = x_2, X_1 = 2$ 的联合概率.

(i) 直接用 (16.23) 式计算条件分布 $f_{X_2|X_1}(x_2|2)$, 并与 (f) 的结果相比较.

(j) 直接用 (16.27) 式求贝叶斯保费, 并和 (g) 的结果相比较.

(k) 确定结构参数 μ, v, a.

(l) 计算 Bühlmann 信度因子和 Bühlmann 信度保费.

(m) 证明: 如果每次有放回地取出 1 个球而不是 2 个, 那么 Bühlmann 信度因子不变.

表 16-6 习题 16.23 的数据

罐子	0	1	2
1	0.40	0.35	0.25
2	0.25	0.10	0.65
3	0.50	0.15	0.35

16.24　假设投保人分为 2 类：A 和 B. 在所有投保人中 A 类占 2/3, B 类占 1/3, 已知每类投保人的年度索赔次数和个体损失额的信息如下：

类型	索赔次数		个体损失额	
	均值	方差	均值	方差
A	0.2	0.2	200	4 000
B	0.7	0.3	100	1 500

某投保人最近 4 年的总索赔额是 500, 求信度因子 Z 以及下一年的信度保费.

16.25　对某个险种, 设 Θ_1 代表索赔次数的风险参数, Θ_2 代表个体索赔额的风险参数, 且 Θ_1 和 Θ_2 相互独立. 还假设在给定 $\Theta_1=\theta_1$ 条件下索赔次数 N 服从 Poisson 分布, 在给定 $\Theta_2=\theta_2$ 条件下个体索赔额 Y 服从指数分布. 已知索赔次数与个体索赔额各自的条件期望和条件方差的期望值, 且索赔次数的条件期望的方差分别是

$$\mu_N=0.1,\quad v_N=0.1,\quad a_N=0.05,\quad \mu_Y=100,\quad v_Y=25\,000.$$

一个投保人最近 3 年的总索赔额是 200, 求 Bühlmann 信度因子和 Bühlmann 信度保费.

16.26　假设在给定 Θ 的条件下 $X_1,\cdots,X_n$ 相互独立, 且有

$$\mathrm{E}(X_j|\Theta)=\beta_j\mu(\Theta)\text{ 和 }\mathrm{Var}(X_j|\Theta)=\tau_j(\Theta)+\psi_j v(\Theta),\quad j=1,\cdots,n.$$

记

$$\mu=\mathrm{E}[\mu(\Theta)],\quad v=\mathrm{E}[v(\Theta)],\quad \tau_j=\mathrm{E}[\tau_j(\Theta)],\quad a=\mathrm{Var}[\mu(\Theta)].$$

(a) 证明

$$\mathrm{E}(X_j)=\beta_j\mu,\quad \mathrm{Var}(X_j)=\tau_j+\psi_j v+\beta_j^2 a,$$

和

$$\mathrm{Cov}(X_i,X_j)=\beta_i\beta_j a,\quad i\neq j.$$

(b) 求解 $\tilde{\alpha}_0,\tilde{\alpha}_1,\cdots,\tilde{\alpha}_n$ 的正交方程组以证明信度保费满足

$$\tilde{\alpha}_0+\sum_{j=1}^{n}\tilde{\alpha}_jX_j=(1-Z)\mathrm{E}(X_{n+1})+Z\beta_{n+1}\bar{X},$$

其中

$$\begin{aligned}m_j&=\beta_j^2(\tau_j+\psi_j v)^{-1},\quad j=1,\cdots,n,\\ m&=m_1+\cdots+m_n,\\ Z&=am(1+am)^{-1},\\ \bar{X}&=\sum_{j=1}^{n}\frac{m_j}{m}\frac{X_j}{\beta_j}.\end{aligned}$$

16.27　对习题 12.72 中的情形求 $\mu(\theta)$ 和贝叶斯保费 $\mathrm{E}(X_{n+1}|\boldsymbol{x})$. 为什么贝叶斯保费等于信度保费?

16.28 对习题 12.73 中的情形求 $\mu(\theta)$ 和贝叶斯保费 $\mathrm{E}(X_{n+1}|\boldsymbol{x})$, 并直接证明信度保费等于贝叶斯保费.

16.29 对习题 12.74 中的情形求 $\mu(\theta)$ 和贝叶斯保费 $\mathrm{E}(X_{n+1}|\boldsymbol{x})$, 并直接证明信度保费等于贝叶斯保费.

16.30 考虑推广的线性指数分布族

$$f(x;\theta,m)=\frac{p(m,x)\mathrm{e}^{-m\theta x}}{[q(\theta)]^m},$$

如果 m 是参数, 则上式称为**离散指数分布族**. 在习题 12.79 中证明了这个随机变量的期望是 $-q'(\theta)/q(\theta)$. 在这里假设 m 已知.

(a) 对先验分布

$$\pi(\theta)=\frac{[q(\theta)]^{-k}\exp(-\theta\mu k)}{c(\mu,k)},\quad \theta_0<\theta<\theta_1 \text{ 满足 } \pi(\theta_0)=\pi(\theta_1),$$

计算贝叶斯保费.

(b) 对同样的先验分布计算 Bühlmann 保费.

(c) 证明逆高斯分布属于离散指数分布族.

16.31 假设在给定 Θ 的条件下 $X_1,\cdots,X_n$ 相互独立, 且有

$$\mathrm{E}(X_j|\Theta)=\tau^j\mu(\Theta) \text{ 且 } \mathrm{Var}(X_j|\Theta)=\frac{\tau^{2j}v(\Theta)}{m_j},\quad j=1,\cdots,n.$$

设

$$\mu=\mathrm{E}[\mu(\Theta)],\ v=\mathrm{E}[v(\Theta)],\ a=\mathrm{Var}[\mu(\Theta)],\ k=v/a,\ \text{且}\ m=m_1+\cdots+m_n.$$

(a) 讨论以上假定在什么情况下是合理的.

(b) 证明

$$\mathrm{E}(X_j)=\tau^j\mu,\quad \mathrm{Var}(X_j)=\tau^{2j}(a+v/m_j),$$

以及

$$\mathrm{Cov}(X_i,X_j)=\tau^{i+j}a,\quad i\neq j.$$

(c) 通过解 $\tilde{\alpha}_0,\tilde{\alpha}_1,\cdots,\tilde{\alpha}_n$ 的正交方程组以证明信度保费满足

$$\tilde{\alpha}_0+\sum_{j=1}^n\tilde{\alpha}_jX_j=\frac{k}{k+m}\tau^{n+1}\mu+\frac{m}{k+m}\sum_{j=1}^n\frac{m_j}{m}\tau^{n+1-j}X_j.$$

(d) 解释 (c) 中的公式.

(e) 假设

$$f_{X_j|\Theta}(x_j|\theta)=\frac{p(x_j,m_j,\tau)\mathrm{e}^{-m_j\tau^{-j}x_j\theta}}{[q(\theta)]^{m_j}}.$$

证明 $\mathrm{E}(X_j|\Theta)=\tau^j\mu(\Theta)$ 和 $\mathrm{Var}(X_j|\Theta)=\tau^{2j}v(\Theta)/m_j$, 其中 $\mu(\theta)=-\frac{\mathrm{d}}{\mathrm{d}\theta}\ln q(\theta)$ 和 $v(\theta)=-\mu'(\theta)$.

(f) 证明如果 Θ 有密度函数

$$\pi(\theta)=\frac{[q(\theta)]^{-k}\mathrm{e}^{-\theta\mu k}}{c(\mu,k)},\quad \theta_0<\theta<\theta_1,$$

其中 $\pi(\theta_0)=\pi(\theta_1)=0$, 则满足精确信度条件.

16.32 给定 $\Theta=\theta$ 的条件下随机变量 $X_1,\cdots,X_n$ 相互独立, 且均服从均值为 θ 的 Poisson 分布, 概率函数是

$$f_{X_j|\Theta}(X_j|\theta)=\frac{\theta^{x_j}\mathrm{e}^{-\theta}}{x_j!},\quad x_j=0,1,2,\cdots.$$

(a) 令 $S=X_1+\cdots+X_n$, 证明 S 的概率密度是

$$f_S(s)=\int_0^{\infty}\frac{(n\theta)^s\mathrm{e}^{-n\theta}}{s!}\pi(\theta)\mathrm{d}\theta,\quad s=0,1,2,\cdots,$$

其中 $\pi(\theta)$ 是 Θ 的概率密度函数.

(b) 证明贝叶斯保费是

$$\mathrm{E}(X_{n+1}|X_1+\cdots+X_n=s)=\frac{s+1}{n}\frac{f_S(s+1)}{f_S(s)},$$

其中 $s=\sum\limits_{j=1}^{n}x_j$.

(c) 如果 (a) 中的 $\pi(\theta)$ 是 gamma 分布, 求 S 的分布函数. 它是哪种类型的分布?

16.33 假设 $X_j|\Theta$ 服从均值为 Θ, 方差为 v 的正态分布, 其中 $\mathrm{j}=1,2,\cdots,n+1$, 并且 Θ 服从均值为 μ, 方差为 a 的正态分布. 也就是有

$$f_{X_j|\Theta}(x_j|\theta)=(2\pi v)^{-1/2}\exp\left[-\frac{1}{2v}(x_j-\theta)^2\right],\quad -\infty<x_j<\infty,$$

及

$$\pi(\theta)=(2\pi a)^{-1/2}\exp\left[-\frac{1}{2a}(\theta-\mu)^2\right],\quad -\infty<\theta<\infty.$$

求后验分布 $\Theta|\boldsymbol{X}$ 和预测分布 $X_{n+1}|\boldsymbol{X}$. 再计算贝叶斯估计量 $\mathrm{E}(X_{n+1}|\boldsymbol{X})$, 同时证明贝叶斯估计等于 Bühlmann 估计.

16.34* 有 2 个罐子, 其中 1 个装有 4 个球, 编号 1~4, 另 1 个有 6 个球, 编号 1~6. 随机挑选 1 个罐, 从中有放回地取出 1 个球, 看到编号是 4. 然后要从同 1 个罐子中再取 1 个球.

(a) 用贝叶斯方法估计下一个球的期望号码.

(b) 用 Bühlmann 信度理论估计下一个球的期望号码.

16.35 一个随机挑选的被保险人的索赔次数服从参数为 θ 的 Poisson 分布, 而在人群中 θ 的概率密度函数是 $\pi(\theta)=3\theta^{-4}$, $\theta>1$. 每个人的参数随着时间的推移保持不变. 在过去两年中某被保险人一共发生了 20 次索赔.

(a)* 计算该被保险人未来期望索赔频率的 Bühlmann 信度估计.

(b) 计算该被保险人未来期望索赔频率的贝叶斯信度估计.

16.36* 某被保险人的索赔分布随时间保持不变, 如果一年半数据的 Bühlmann 信度因子是 0.5, 求三年数据的 Bühlmann 信度因子.

16.37* 3 个罐子中含有编号是 0 或 1 的球. A 罐中 10%的球编号是 0, B 罐中 60%的球编号是 0, C 罐中 80%的球编号是 0. 先随机挑选 1 个罐, 然后从中有放回地取出 3 个球, 已知这 3 个球编号总和是 1. 在同一个罐中再次有放回地取出 3 个球.

(a) 用贝叶斯公式计算第 2 次取出的 3 个球编号总和的期望值.

(b) 用 Bühlmann 信度公式计算第 2 次取出的 3 个球编号总和的期望值.

16.38* 设索赔次数服从参数为 λ 的 Poisson 分布. 某被保险人过去 3 年有 3 次索赔.

(a) λ 的概率密度是 $f(\lambda)=4\lambda^{-5}$, $\lambda>1$. 求 Bühlmann 信度公式中的 K 值, 并用 Bühlmann 信度估计该被保险人的索赔频率.

(b) λ 的概率密度是 $f(\lambda)=1, 0<\lambda<1$. 求 Bühlmann 信度公式中的 K 值, 并用 Bühlmann 信度估计该被保险人的索赔频率.

16.39* 设索赔次数服从参数为 h 的 Poisson 分布. h 服从密度函数是 $f(h)=he^{-h}, h>0$ 的 gamma 分布. 确定单个观察值对应的 Bühlmann 信度因子. (贝叶斯信度已在习题 12.86 解决)

16.40 考虑习题 12.88 的情形.

(a) 用贝叶斯信度计算第二年的期望索赔次数.

(b)* 用 Bühlmann 信度计算第二年的期望索赔次数.

16.41* 有 3 个转盘, 每个转盘被分成相同大小的 6 块区域, 上面标有数字 0, 12, 48. 每个数字在不同的转盘上分别占有的区域数如下. 转盘 A: 2, 2, 2; 转盘 B: 3, 2, 1; 转盘 C: 4, 1, 1. 从这 3 个转盘中随机选出 1 个, 已知在第 1 次旋转中得到 0.

(a) 用 Bühlmann 信度估计同一个转盘第二次旋转所得的期望值.

(b) 用贝叶斯信度估计同一个转盘第二次旋转所得的期望值.

16.42 已知每年的索赔次数服从均值为 λ 的 Poisson 分布, 而参数 λ 服从 $(1,3)$ 上的均匀分布.

(a)* 如果随机挑选出的一个人没有发生索赔, 求该事件的概率.

(b)* 如果一个被保险人在第一年中有一次索赔, 用 Bühlmann 信度估计他第二年的期望索赔次数.

(c) 如果一个被保险人在第一年中有一次索赔, 用贝叶斯信度估计他第二年的期望索赔次数.

16.43* 2 个团体 A 和 B 含有相同的风险量. 团体 A 中单位风险每年的索赔次数的均值是 1/6, 方差是 5/36; 个体索赔量的均值是 4, 方差是 20. 团体 B 中单位风险每年的索赔次数的均值是 5/6, 方差是 5/36; 个体索赔量的均值是 2, 方差是 5. 随机从这 2 个团体中选取 1 个单位风险, 并连续观察其 4 年的数据.

(a) 用 Bühlmann 信度计算净保费的信度因子 Z.

(b) 假设 4 年的年平均损失量是 0.25, 求该风险净保费的 Bühlmann 信度估计.

16.44* 设 X_1 是单次试验的结果, $\mathrm{E}(X_2|X_1)$ 是第二次试验的期望值. 给出如下信息:

结果 T	$\Pr(X_1=T)$	$\mathrm{E}(X_2\|X_1=T)$ 的 Bühlmann 估计	$\mathrm{E}(X_2\|X_1=T)$ 的贝叶斯估计
1	1/3	2.72	2.6
8	1/3	7.71	7.8
12	1/3	10.57	—

求 $\mathrm{E}(X_2|X_1=12)$ 的贝叶斯估计.

16.45 考虑习题 12.90 的情形.

(a) 用贝叶斯信度计算第二年的期望索赔次数.

(b)* 用 Bühlmann 信度计算第二年的期望索赔次数.

16.46 考虑习题 12.91 的情形.

(a) 用贝叶斯信度计算第二年的期望索赔次数.

(b) 用 Bühlmann 信度计算第二年的期望索赔次数.

16.47 用 2 个转盘 A_1 和 A_2 来确定索赔次数. 转盘 A_1 有 0.15 的概率转到 1 次索赔, 0.85 的概率无索赔. 转盘 A_2 有 0.05 的概率转到 1 次索赔, 0.95 的概率无索赔. 如果发生索赔, 则另外 2 个转盘 B_1 和 B_2 将用于确定索赔量. B_1 有 0.8 的概率产生索赔量 20, 0.2 的概率产生索赔量 40. B_2 有 0.3 的概率产生索赔量 20, 0.7 的概率产生索赔量 40. 从 A_1 和 A_2, B_1 和 B_2 中分别选取 1 个转盘, 对这 2 个转盘进行 3 次操作, 得到索赔量 0, 20, 0.

(a)* 用 Bühlmann 信度理论分别估计期望索赔次数和期望个体索赔量, 并用这些估计值计算对同一对转盘下一次的期望观察值.

(b) 对以上 3 个观察值, 只用一次 Bühlmann 信度公式估计这一对转盘下一次操作的期望值.

(c)* 利用贝叶斯估计解答 (a) 和 (b).

(d)* 对同一对转盘, 求

$$\lim_{n\to\infty}\mathrm{E}(X_n|X_1=X_2=\cdots=X_{n-1}=0).$$

16.48* 一个风险组合中所有风险都服从正态分布. A 类风险均值是 0.1, 标准差是 0.03; B 类风险均值是 0.5, 标准差是 0.05; C 类风险均值是 0.9, 标准差是 0.01. 每类风险的风险量相同. 现有某个风险的观察值是 0.12, 求这个风险下一个观察值的贝叶斯期望.

16.49* 已知如下信息.

(1) 条件分布 $f_{X|\Theta}(x|\theta)$ 属于线性指数分布族.

(2) 先验分布 $\pi(\theta)$ 是 $f_{X|\Theta}(x|\theta)$ 的共轭先验分布.

(3) $\mathrm{E}(X)=1$.

(4) $\mathrm{E}(X|X_1=4)=2$, 其中 X_1 是单个观察值.

(5) 条件方差的期望 $\mathrm{E}[\mathrm{Var}(X|\Theta)]=3$.

计算条件期望的方差 $\mathrm{Var}[\mathrm{E}(X|\Theta)]$.

16.50* 已知如下信息.

(1) X 是均值为 μ, 方差为 v 的随机变量.

(2) μ 是均值为 2, 方差为 4 的随机变量.

(3) v 是均值为 8, 方差为 32 的随机变量.

根据 X 的 3 个观察值计算 Bühlmann 信度因子 Z.

16.51 个体索赔额的密度函数是 $f_{Y|\Lambda}(y|\lambda) = \lambda^{-1}e^{-y/\lambda}, y, \lambda > 0$. 参数 λ 服从逆 gamma 分布, 密度函数是 $\pi(\lambda) = 400\lambda^{-3}e^{-20/\lambda}$.

(a)* 计算无条件期望 E(X).

(b) 设观察到的两次索赔分别是 15 和 25, 对同一个被保险人用 Bühlmann 信度计算下一次索赔额的期望值.

(c) 用贝叶斯信度估计求解 (b).

16.52 索赔次数服从 $n = 1$, θ 未知的二项分布. 参数 θ 的均值是 0.25, 方差是 0.07. 对单个观察值计算 Bühlmann 信度因子 Z.

16.53* 4 个射手各自朝相距 100 英尺的目标射击, 相邻目标之间相隔 2 英尺 (也就是 4 个目标位于 1 条直线上, 离原点分别是 0, 2, 4, 6 英尺). 射手射击时只会向左或向右偏, 而不会向上或向下偏. 每个射手的射击方向服从正态分布, 均值就是他的目标, 标准差是离目标距离的固定常数倍. 100 英尺时的标准差是 3 英尺. 通过观察某个射手在直线上的射击点, 可以估计该射手下一枪的大致位置.

(a) 对一个随机挑选的射手的单个观察值, 求 Bühlmann 信度因子.

(b) 哪种情形能得到最大的 Bühlmann 信度因子?

i. 把目标变成 0, 4, 8, 12.

ii. 把射手们移至距离目标 60 英尺.

iii. 把目标变成 2, 2, 10, 10.

iv. 把对同一个射手观察的次数由 1 增加到 3.

v. 把 2 个射手移至距离目标 50 尺, 并把观察次数增加到 2.

16.54* 风险 1 分别以 0.5, 0.3, 0.2 的概率产生索赔量 100, 1 000, 20 000. 对风险 2, 以上索赔量的概率分别是 0.7, 0.2, 0.1. 风险 1 发生的概率是风险 2 的 2 倍, 现在观察到一笔 100 的索赔, 但不确定来自哪个风险.

(a) 用贝叶斯信度计算同一风险的第二笔索赔的期望值.

(b) 用 Bühlmann 信度计算同一风险的第二笔索赔的期望值.

16.55* 已知如下信息.

(1) 单个被保险人的索赔次数服从均值为 M 的 Poisson 分布.

(2) 个体索赔额服从指数分布, 密度函数是 $f_{X|\Lambda}(x|\lambda) = \lambda^{-1}e^{-x/\lambda}, x, \lambda > 0$

(3) M 和 Λ 独立.

(4) $\mathrm{E}(M) = 0.10$, $\mathrm{Var}(M) = 0.002\ 5$.

(5) $\mathrm{E}(\Lambda) = 1\ 000$, $\mathrm{Var}(\Lambda) = 640\ 000$.

(6) 索赔次数和个体索赔额独立.

(a) 计算单个被保险人净保费的条件方差的期望.

(b) 计算单个被保险人净保费的条件期望的方差.

16.56 在例 16.24 中, 如果 $\rho=0$, 则 $Z=0$, μ 就是估计量. 也就是说, 经验数据可以被忽略. 然而当 ρ 逐渐增加趋近于 1 时, Z 也逐渐增加至 1, 结果是样本均值对预测 X_{n+1} 将更有价值. 解释该结果的合理性.

16.57 这一题要求读者推导例 16.33 中的一些细节.

(a) **digamma** 函数的正式定义是 $\psi(\alpha)=\Gamma'(\alpha)/\Gamma(\alpha)$, 从定义证明

$$\Psi(\alpha)=\frac{1}{\Gamma(\alpha)}\int_0^\infty(\ln x)x^{\alpha-1}\mathrm{e}^{-x}\mathrm{d}x.$$

(b) **trigamma** 函数的正式定义是 $\psi'(\alpha)$. 用 trigamma 函数, digamma 函数和 gamma 函数来表示以下表达式

$$\int_0^\infty(\ln x)^2x^{\alpha-1}\mathrm{e}^{-x}\mathrm{d}x.$$

16.58 考虑一个与例 16.32 和例 16.33 相似的情形. 个体观测样本的大小是 25, 来自 μ 未知, $\sigma=2$ 的对数正态分布. Θ 的先验分布 (Θ 代表未知的 μ) 是均值为 5, 标准差为 1 的正态分布. 分别计算贝叶斯估计量, 信度估计量和对数信度估计量, 并比较它们的均方误差. 采用例 16.32 和例 16.33 中的分位数进行计算.

16.59 用随机向量 $\boldsymbol{X}$ 表示过去的所有数据, X_{n+1} 代表下一个观察值. $g(\boldsymbol{X})$ 可以是过去数据的任意函数.

(a) 证明

$$\begin{aligned}\mathrm{E}\{[X_{n+1}-g(\boldsymbol{X})]^2\}=&\mathrm{E}\{[X_{n+1}-\mathrm{E}(X_{n+1}|\boldsymbol{X})]^2\}\\&+\mathrm{E}\{[\mathrm{E}(X_{n+1}|\boldsymbol{X})-g(X)]^2\},\end{aligned}$$

其中的期望是对 $(X_{n+1},\boldsymbol{X})$ 进行的计算.

(b) 证明: 当 $g(\boldsymbol{X})$ 等于贝叶斯保费 (预测分布的期望值) 时, 期望平方误差 $\mathrm{E}\{[X_{n+1}-g(\boldsymbol{X})]^2\}$ 取到最小值.

(c) 证明: 如果把 $g(\boldsymbol{X})$ 限定在过去数据的线性函数上, 那么当它等于信度保费时, 期望平方误差 $\mathrm{E}\{[X_{n+1}-g(\boldsymbol{X})]^2\}$ 取值最小.

16.5 经验贝叶斯参数估计

在 16.4 节中, 为了能把经验数据用于厘定预期费率, 提出了一种建模方法, 并得到贝叶斯保费和信度保费作为参考值. 但是在实际应用中, 还有需要解决的问题.

前面的例子之所以能够计算出要求的量, 是因为假设分布 $f_{X_j|\Theta}(x_j|\theta)$ 和 $\pi(\theta)$ 均为已知. 尽管这些例子有助于我们理解方法本身, 却很难精确地代表保险组合的

情形. 在现实中运用模型时需要引入参数, 并通过恰当选取参数以使模型与现实较好地吻合. 这方面的例子有: 适当选取 Poisson-gamma 模型 (例 16.15) 中 gamma 分布的参数 α 和 β; 或者适当选取 Bühlmann 或 Bühlmann-Straub 模型中的参数 μ, v 和 a. 要想具体计算出贝叶斯保费或是信度保费, 需要把当中的参数换成具体的数值.

一般说来, 未知参数会与结构密度 $\pi(\theta)$ 相关, 因此把这些参数称为**结构参数**. 这里使用的术语从属于前述贝叶斯理论的框架. 严格地说, 在贝叶斯情形中认为所有结构参数都是已知的, 不需要进行估计. 一个例子是 Poisson-gamma 分布中关于结构密度的先验信息通过选取 $\alpha = 360, \beta = 1/240$ 进行了量化. 考虑到最终目的, 这种完全信息的贝叶斯方法经常不能让人满意 (e.g. 当只有很少甚至没有先验信息可用, 比如说要考虑一种新的险种), 这时就需要利用手边的数据估计结构 (先验) 参数. 这种方法就叫做**经验贝叶斯估计**.

当对 $\pi(\theta)$ 和 $f_{X_j|\Theta}(x_j|\theta)$ 的信息几乎一无所知时 (如, 在 Bühlmann 或 Bühlmann-Straub 模型中只需要知道前两阶矩), 称之为非参数情形, 将在 16.5.1 节中讨论. 如果假设 $f_{X_j|\Theta}(x_j|\theta)$ 具有某种带参数的形式 (例如 Poisson 分布或正态分布), 而 $\pi(\theta)$ 并非如此, 则称为半参数情形, 将在 16.5.2 节中讨论. 最后当 $f_{X_j|\Theta}(x_j|\theta)$ 和 $\pi(\theta)$ 都具有某种带参数的特定形式时, 称为完全参数情形 (在技术处理上将更复杂), 将在 16.5.3 节中简要讲述.

选取参数模型还是非参数模型, 部分取决于现实中的具体情形, 另外还取决于分析师的知识水平和判断能力. 例如, 在分析索赔次数时可能会假设 $f_{X_j|\Theta}(x_j|\theta)$ 服从 Poisson 分布, 但要假设 $\pi(\theta)$ 也具有参数模型就有些不够合理.

任何关于参数的假设都应当 (尽可能地) 在参数估计中反映出来. 比如, 在 Poisson 分布的假设下均值和方差相等, 因此应当使用同一个数作为它们的估计. 在正常情况下, 如果模型本身是合理的, 非参数估计量会比参数估计量更有效. 在决定是否选择参数模型时应当考虑这一点.

最后, 非参数模型的优点是能处理广泛的情形, 使得关于参数的各种假设有时显得多余 (这些假设常常会具有某些不合理之处).

本节中数据将以如下形式表示: 对 $r \geqslant 1$ 个投保人, 当中的每个人的单位风险损失量是 $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{in_i})^{\mathrm{T}}$, $i = 1, \cdots, r$. 随机向量 $\{\boldsymbol{X}_i, i = 1, \cdots, r\}$ 在统计上认为是相互独立的 (即假设不同投保人的经验数据相互独立). 第 i 个投保人的未知风险参数是 θ_i, $i = 1, \cdots, r$, 并进一步假定 θ_i 是独立同分布的, 具有结构密度 $\pi(\theta_i)$ 的随机变量 Θ_i 的一次实现 ($i = 1, \cdots, r$). 对固定的 i, 假设条件随机变量 $X_{ij}|\Theta_i$ 相互独立, 有概率密度 $f_{X_{ij}|\Theta}(x_{ij}|\theta)$, $j = 1, \cdots, n_i$.

有两个常见的情形可以产生上述的数据形式. 第一种情形是分类费率厘定. 下标 i 指代类别或团体, j 指代当中的个体. 第二种情形类似, 下标 i 仍指代类别或

团体, j 代表年份, 以每年的平均损失量作为观察值. 关于第二种情形的一个例子见 Meyers[91], 其中观察 $j=1,2,3$ 年的数据有 $i=1,\cdots,319$ 种职业分类. 下面统一指代 r 个实体为 "投保人" .

已知数据还应当包括投保人 i 的风险量向量 $\boldsymbol{m}_i=(m_{i1},\cdots,m_{in_i})^T, i=1,\cdots,r$. 如果不知道这个风险量向量, (只要合适的话) 可以令 $m_{ij}=1$ 对所有的 i,j 均成立. 为了表达上的方便, 记

$$m_i=\sum_{j=1}^{n_i}m_{ij},\quad i=1,\cdots,r,$$

为投保人 i 过去的总风险量, 并记

$$\bar{X}_i=\frac{1}{m_i}\sum_{j=1}^{n_i}m_{ij}X_{ij},\quad i=1,\cdots,r.$$

为过去单位风险的平均损失量. 另外, 所有投保人总风险量是

$$m=\sum_{i=1}^{r}m_i=\sum_{i=1}^{r}\sum_{j=1}^{n_i}m_{ij}.$$

总的单位风险平均损失是

$$\bar{X}=\frac{1}{m}\sum_{i=1}^{r}m_i\bar{X}_i=\frac{1}{m}\sum_{i=1}^{r}\sum_{j=1}^{n_i}m_{ij}X_{ij}. \tag{16.54}$$

需要估计的参数依赖于对分布 $f_{X_{ij}|\Theta}(x_{ij}|\theta)$ 和 $\pi(\theta)$ 的假定.

对 Bühlmann-Straub 公式还要考虑另外一些量. 设条件期望 (与 j 无关) 是

$$\mathrm{E}(X_{ij}|\Theta_i=\theta_i)=\mu(\theta_i).$$

条件方差是

$$\mathrm{Var}(X_{ij}|\Theta_i=\theta_i)=\frac{v(\theta_i)}{m_{ij}}.$$

结构参数是

$$\mu=\mathrm{E}[\mu(\Theta_i)],\quad v=\mathrm{E}[v(\Theta_i)],$$

和

$$a=\mathrm{Var}[\mu(\Theta_i)].$$

本节将给出利用数据估计未知的 μ,v 和 a 的方法. 投保人 i 下一年单位风险的信度保费是

$$Z_i\bar{X}_i+(1-Z_i)\mu,\quad i=1,\cdots,r, \tag{16.55}$$

其中

$$Z_i = \frac{m_i}{m_i + k}, \quad k = \frac{v}{a}.$$

用 $\hat{\mu}$, $\hat{v}$ 和 $\hat{a}$ 分别代表 μ, v 和 a 的估计量, 则可以把信度保费 (16.55) 换成它的估计量

$$\hat{Z}_i \bar{X}_i + (1 - \hat{Z}_i)\hat{\mu}, \tag{16.56}$$

其中

$$\hat{Z}_i = \frac{m_i}{m_i + \hat{k}}, \quad \hat{k} = \frac{\hat{v}}{\hat{a}}.$$

注意到即使 $\hat{v}$ 和 $\hat{a}$ 都是 v 和 a 的无偏估计量, $\hat{k}$ 和 $\hat{Z}_i$ 也未必是 k 和 Z_i 的无偏估计量. 最后, 对投保人 i 下一年度 m_{i,n_i+1} 风险量的信度保费就是 (16.56) 式乘以 m_{i,n_i+1}.

16.5.1 非参数估计

本节讨论 μ, v 和 a 的无偏估计量. 为了阐述估计的想法, 下面首先介绍一些简单的 Bühlmann 模型的例子.

例 16.34 *假设对所有的 i 有 $n_i = n > 1$, 对所有的 i 和 j 有 $m_{ij} = 1$. 也就是投保人 i 的损失量向量是*

$$\boldsymbol{X}_i = (X_{i1}, \cdots, X_{in})^{\mathrm{T}}, \quad i = 1, \cdots, r.$$

另外, 在给定 $\Theta_i = \theta_i$ 的条件下, X_{ij} 的均值是

$$\mu(\theta_i) = \mathrm{E}(X_{ij}|\Theta_i = \theta_i),$$

方差是

$$v(\theta_i) = \mathrm{Var}(X_{ij}|\Theta_i = \theta_i),$$

且 $X_{i1}, \cdots, X_{in}$ 条件独立. 不同投保人的历史数据相互独立, 因此如果 $i \neq s$, 则 X_{ij} 与 X_{st} 独立. 在这个例子中

$$\bar{X}_i = n^{-1} \sum_{j=1}^{n} X_{ij} \text{ and } \bar{X} = r^{-1} \sum_{i=1}^{r} \bar{X}_i = (rn)^{-1} \sum_{i=1}^{r} \sum_{j=1}^{n} X_{ij}.$$

求 Bühlmann 模型相关量的无偏估计.

解 μ 的一个无偏估计量是

$$\hat{\mu} = \bar{X}.$$

这是因为

$$\begin{aligned}\mathrm{E}(\hat{\mu}) &= (rn)^{-1}\sum_{i=1}^{r}\sum_{j=1}^{n}\mathrm{E}(X_{ij}) = (rn)^{-1}\sum_{i=1}^{r}\sum_{j=1}^{n}\mathrm{E}[\mathrm{E}(X_{ij}|\Theta_i)]\\ &= (rn)^{-1}\sum_{i=1}^{r}\sum_{j=1}^{n}\mathrm{E}[\mu(\Theta_i)] = (rn)^{-1}\sum_{i=1}^{r}\sum_{j=1}^{n}\mu = \mu.\end{aligned}$$

要估计 v, 考虑

$$\hat{v}_i = \frac{1}{n-1}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)^2.$$

由于对固定的 i, 随机变量 $X_{i1},\cdots,X_{in}$ 在给定 $\Theta_i=\theta_i$ 的条件下独立, 所以 $\hat{v}_i$ 是 $\mathrm{Var}(X_{ij}|\Theta_i=\theta_i)=v(\theta_i)$ 的无偏估计. 在无条件下有

$$\mathrm{E}(\hat{v}_i) = \mathrm{E}[\mathrm{E}(\hat{v}_i|\Theta_i)] = \mathrm{E}[v(\Theta_i)] = v,$$

故 $\hat{v}_i$ 是 v 的无偏估计. 于是可得 v 的一个无偏估计是

$$\hat{v} = \frac{1}{r}\sum_{i=1}^{r}\hat{v}_i = \frac{1}{r(n-1)}\sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)^2. \tag{16.57}$$

接下来估计参数 a. 首先有

$$\mathrm{E}(\bar{X}_i|\Theta_i=\theta_i) = n^{-1}\sum_{j=1}^{n}\mathrm{E}(X_{ij}|\Theta_i=\theta_i) = n^{-1}\sum_{j=1}^{n}\mu(\theta_i) = \mu(\theta_i).$$

因此可得

$$\mathrm{E}(\bar{X}_i) = \mathrm{E}[\mathrm{E}(\bar{X}_i|\Theta_i)] = \mathrm{E}[\mu(\Theta_i)] = \mu,$$

以及

$$\begin{aligned}\mathrm{Var}(\bar{X}_i) &= \mathrm{Var}[\mathrm{E}(\bar{X}_i|\Theta_i)] + \mathrm{E}[\mathrm{Var}(\bar{X}_i|\Theta_i)]\\ &= \mathrm{Var}[\mu(\Theta_i)] + \mathrm{E}\left[\frac{v(\Theta_i)}{n}\right] = a + \frac{v}{n}.\end{aligned}$$

由此可知, $\bar{X}_1,\cdots,\bar{X}_r$ 独立, 且有共同的期望 μ 和方差 $a+v/n$. 它们的样本均值是 $\bar{X}=r^{-1}\sum\limits_{i=1}^{r}\bar{X}_i$, 从而 $a+v/n$ 的一个无偏估计是 $(r-1)^{-1}\sum\limits_{i=1}^{r}(\bar{X}_i-\bar{X})^2$. 前面已经得到 v 的一个无偏估计, 所以 a 的一个无偏估计是

$$\hat{a} = \frac{1}{r-1}\sum_{i=1}^{r}(\bar{X}_i-\bar{X})^2 - \frac{\hat{v}}{n}$$

$$= \frac{1}{r-1}\sum_{i=1}^{r}(\bar{X}_i - \bar{X})^2 - \frac{1}{rn(n-1)}\sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2. \qquad (16.58)$$

□

这些估计量似曾相识. 考虑在单因素方差分析中, 每个投保人代表一种处理方法. v 的估计量 (16.57) 就是组内 (也叫误差) 平方和的平均, 而 a 的估计量 (16.58) 的第一项就是组间 (也叫不同的处理方法) 差的平方和除以 n. 当组间平方和相对组内平方和而言比较小时 —— 也就是 $\hat{a}$ 相对 $\hat{v}$ 比较小时, 接受所有处理方法有相同均值的原假设. 这也意味着 $\hat{Z}$ 接近于 0, 即给予每个 $\bar{X}_i$ 很小的信度, 这在所有投保人风险同质的情形下是一个很自然的结论.

由于在 (16.58) 式中出现了减号, 因此可能算出 $\hat{a}$ 是负值. 如果是这样, 习惯上令 $\hat{a} = \hat{Z} = 0$. 这种情形等价于在方差分析中 F 检验统计量小于 1, 它必然导致接受均值相等的原假设.

例 16.35　(例 16.34 续) 作为一个具体的例子, 假设有 $r = 2$ 个投保人, 每人有 $n = 3$ 年的数据. 损失量分别是 $\boldsymbol{x}_1 = (3,5,7)^{\mathrm{T}}$ 和 $\boldsymbol{x}_2 = (6,12,9)^{\mathrm{T}}$. 对每个投保人求 Bühlmann 信度保费.

解　由题设有

$$\bar{X}_1 = \frac{1}{3}(3+5+7) = 5, \quad \bar{X}_2 = \frac{1}{3}(6+12+9) = 9,$$

因此 $\bar{X} = \dfrac{5+9}{2} = 7$. 从而 $\hat{\mu} = 7$. 接下来有

$$\hat{v}_1 = \frac{1}{2}[(3-5)^2 + (5-5)^2 + (7-5)^2] = 4,$$

$$\hat{v}_2 = \frac{1}{2}[(6-9)^2 + (12-9)^2 + (9-9)^2] = 9,$$

所以有 $\hat{v} = \dfrac{1}{2}(4+9) = \dfrac{13}{2}$, 进而有

$$\hat{a} = [(5-7)^2 + (9-7)^2] - \frac{1}{3}\hat{v} = \frac{35}{6}.$$

从而得到$\hat{k} = \dfrac{\hat{v}}{\hat{a}} = \dfrac{39}{35}$, 信度因子的估计是 $\hat{Z} = 3/(3+\hat{k}) = \dfrac{35}{48}$. 对投保人 1 和投保人 2, 信度保费分别是

$$\hat{Z}\bar{X}_1 + (1-\hat{Z})\hat{\mu} = \left(\frac{35}{48}\right)(5) + \left(\frac{13}{48}\right)(7) = \frac{133}{24},$$

$$\hat{Z}\bar{X}_2 + (1-\hat{Z})\hat{\mu} = \left(\frac{35}{48}\right)(9) + \left(\frac{13}{48}\right)(7) = \frac{203}{24}.$$

□

现在考虑更一般的 Bühlmann-Straub 模型假定. 因为 $\mathrm{E}(X_{ij}) = \mathrm{E}[\mathrm{E}(X_{ij}|\Theta_i)] = \mathrm{E}[\mu(\Theta_i)] = \mu$, 所以有

$$\mathrm{E}(\bar{X}_i|\Theta_i) = \sum_{j=1}^{n_i} \frac{m_{ij}}{m_i}\mathrm{E}(X_{ij}|\Theta_i) = \sum_{j=1}^{n_i} \frac{m_{ij}}{m_i}\mu(\Theta_i) = \mu(\theta_i),$$

故

$$\mathrm{E}(\bar{X}_i) = \mathrm{E}[\mathrm{E}(\bar{X}_i|\Theta_i)] = \mathrm{E}[\mu(\Theta_i)] = \mu.$$

最后可得

$$\mathrm{E}(\bar{X}) = \frac{1}{m}\sum_{i=1}^{r} m_i\mathrm{E}(\bar{X}_i) = \frac{1}{m}\sum_{i=1}^{r} m_i\mu = \mu,$$

显然 μ 的一个无偏估计量是

$$\hat{\mu} = \bar{X}. \tag{16.59}$$

现在已知 $\mathrm{E}(X_{ij}|\Theta_i) = \mu(\Theta_i)$ 和 $\mathrm{Var}(X_{ij}|\Theta_i) = v(\Theta_i)/m_{ij}$, $j = 1, \cdots, n_i$, 考虑

$$\hat{v}_i = \frac{\sum_{j=1}^{n_i} m_{ij}(X_{ij} - \bar{X}_i)^2}{n_i - 1}, \quad i = 1, \cdots, r. \tag{16.60}$$

在给定 Θ_i 的条件下, 在 (16.12) 中令 $\beta = 0, \alpha = v(\Theta_i)$ 即得到 $\mathrm{E}(\hat{v}_i|\Theta_i) = v(\Theta_i)$, 这也就是说在无条件情形下有

$$\mathrm{E}(\hat{v}_i) = \mathrm{E}[\mathrm{E}(\hat{v}_i|\Theta_i)] = \mathrm{E}[v(\Theta_i)] = v.$$

因此对 $i = 1, \cdots, r$, $\hat{v}_i$ 是 v 的无偏估计. v 的另一个无偏估计量是上述估计的加权平均 $\hat{v} = \sum\limits_{i=1}^{r} w_i\hat{v}_i$, 其中 $\sum\limits_{i=1}^{r} w_i = 1$. 如果选择权重与 $n_i - 1$ 成比例, 就是让原始的诸 X_{ij} 的权重与 m_{ij} 成比例, 即取 $w_i = (n_i - 1)/\sum\limits_{i=1}^{r}(n_i - 1)$, 可以得到 v 的一个无偏估计量

$$\hat{v} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i} m_{ij}(X_{ij} - \bar{X}_i)^2}{\sum_{i=1}^{r}(n_i - 1)}. \tag{16.61}$$

接下来估计 a. 对固定的 i, 随机变量 $X_{i1}, \cdots, X_{in_i}$ 在给定 Θ_i 的条件下相互独立, 因此有

$$\begin{aligned}\mathrm{Var}(\bar{X}_i|\Theta_i) &= \sum_{j=1}^{n_i}\left(\frac{m_{ij}}{m_i}\right)^2 \mathrm{Var}(X_{ij}|\Theta_i) = \sum_{j=1}^{n_i}\left(\frac{m_{ij}}{m_i}\right)^2 \frac{v(\Theta_i)}{m_{ij}} \\ &= \frac{v(\Theta_i)}{m_i^2}\sum_{j=1}^{n_i} m_{ij} = \frac{v(\Theta_i)}{m_i}.\end{aligned}$$

这也就是说在无条件情形下

$$\begin{aligned}\mathrm{Var}(\bar{X}_i) &= \mathrm{Var}[\mathrm{E}(\bar{X}_i|\Theta_i)] + \mathrm{E}[\mathrm{Var}(\bar{X}_i|\Theta_i)] \\ &= \mathrm{Var}[\mu(\Theta_i)] + \mathrm{E}\left[\frac{v(\Theta_i)}{m_i}\right] = a + \frac{v}{m_i}.\end{aligned} \tag{16.62}$$

概括地说, $\bar{X}_1, \cdots, \bar{X}_r$ 相互独立, 并且有相同的均值 μ, 方差 $\mathrm{Var}(\bar{X}_i) = a + v/m_i$, 还有 $\bar{X} = m^{-1}\sum_{i=1}^{r} m_i\bar{X}_i$. 在 (16.12) 中令 $\beta = a, \alpha = v$ 得到

$$\mathrm{E}\left[\sum_{i=1}^{r} m_i(\bar{X}_i - \bar{X})^2\right] = a\left(m - m^{-1}\sum_{i=1}^{r} m_i^2\right) + v(r-1).$$

把 v 换成 $\hat{v}$, 然后解出 a, 得到 a 的一个无偏估计量是

$$\hat{a} = \left(m - m^{-1}\sum_{i=1}^{r} m_i^2\right)^{-1}\left[\sum_{i=1}^{r} m_i(\bar{X}_i - \bar{X})^2 - \hat{v}(r-1)\right], \tag{16.63}$$

其中 $\hat{v}$ 由 (16.61) 给出. (16.63) 的另一种形式则在习题 16.67 中给出.

这里做一些简单的评述. 等式 (16.59), (16.61) 和 (16.63) 分别给出了 μ, v 和 a 的无偏估计. 它们没有参数形式, 不需要关于分布函数的假设. 当然它们也不是唯一的无偏估计量. 计算中还可能会出现 $\hat{a} < 0$, 这时 a 很有可能非常接近于 0, 因此取 $\hat{Z} = 0$ 是合理的. 此外, 例 16.34 的 Bühlmann 估计量是 $m_{ij} = 1, n_i = n$ 的特殊情形. 还可以从例 16.41 中看到, 当 $X_{ij}|\Theta_i$ 和 Θ_i 都服从正态分布时, 这 3 个估计量实际上也是最大似然估计量, 因此这些估计量会有某些好的统计性质.

关于上述公式的使用还有一个问题. 过去第 i 个投保人的索赔数据来自 m_i 风险量, 所有投保人的总损失是 $TL = \sum_{i=1}^{r} m_i\bar{X}_i$. 如果在过去也按照上面计算出的信度保费收取保费, 则总保费是

$$\begin{aligned}TP &= \sum_{i=1}^{r} m_i[\hat{Z}_i\bar{X}_i + (1-\hat{Z}_i)\hat{\mu}] \\ &= \sum_{i=1}^{r} m_i(1-\hat{Z}_i)(\hat{\mu} - \bar{X}_i) + \sum_{i=1}^{r} m_i\bar{X}i \\ &= \sum_{i=1}^{r} m_i\frac{\hat{k}}{m_i + \hat{k}}(\hat{\mu} - \bar{X}_i) + \sum_{i=1}^{r} m_i\bar{X}_i.\end{aligned}$$

理想情况是 TL 等于 TP. 因为任何增加保费的行为要想得到监管机构的同意, 必须要以过去总的索赔水平为依据. 信度保费有着理论和现实意义, 同时如果还能够保持总保费收入与总损失相匹配就更好了. 这样需要有

$$0 = \sum_{i=1}^{r} m_i\frac{\hat{k}}{m_i + \hat{k}}(\hat{\mu} - \bar{X}_i),$$

也就是

$$\hat{\mu}\sum_{i=1}^{r}\hat{Z}_i=\sum_{i=1}^{r}\hat{Z}_i\bar{X}_i,$$

或是

$$\hat{\mu}=\frac{\sum_{i=1}^{r}\hat{Z}_i\bar{X}_i}{\sum_{i=1}^{r}\hat{Z}_i}. \tag{16.64}$$

这表示应该使用个体样本均值的信度加权平均作为 $\hat{\mu}$, 而不是用 (16.59). 这两种算法都是无偏估计 (给定诸 $\hat{Z}_i$), 但是前者的优点是保持总损失不变. 利用最小二乘法也能推导该结果. 注意当使用 (16.63) 时, 仍使用从 (16.54) 得到的 $\bar{X}$. 最后, 由例 16.7 以及 (16.62) 中 $\mathrm{Var}(\bar{X}_j)$ 的形式, 可知 (16.64) 的权重能使 $\hat{\mu}$ 取到最小的无条件方差.

例 16.36 两组投保人的经验数据在表 16-7 给出. 估计第 4 年每组分别应当收取的信度保费.

表 16-7 例 16.36 的数据

	投保人	第 1 年	第 2 年	第 3 年	第 4 年
总索赔数	1	—	10 000	13 000	—
团体中人数		—	50	60	75
总索赔数	2	18 000	21 000	17 000	—
团体中人数		100	110	105	90

解 首先要计算各组每人的平均索赔量. 对第一组有 $n_1=2$, 对第二组有 $n_2=3$. 对投保人 1 而言, 经验数据具体是哪一年的并不重要, 所以为了表示方便, 令

$$m_{11}=50,\ X_{11}=\frac{10\ 000}{50}=200.$$

类似地有

$$m_{12}=60,\ X_{12}=\frac{13\ 000}{60}=216.67.$$

因此

$$m_1=m_{11}+m_{12}=50+60=110,$$

$$\bar{X}_1=\frac{10\ 000+13\ 000}{110}=209.09.$$

对投保人 2 有

$$m_{21}=100,X_{21}=\frac{18\ 000}{100}=180,$$

$$m_{22}=110,X_{22}=\frac{21\ 000}{110}=190.91,$$

$$m_{23}=105,X_{23}=\frac{17\ 000}{105}=161.90.$$

故

$$m_2 = m_{21} + m_{22} + m_{23} = 100 + 110 = 105 = 315,$$

$$\bar{X}_2 = \frac{18\ 000 + 21\ 000 + 17\ 000}{315} = 177.78.$$

并且, $m = m_1 + m_2 = 110 + 315 = 425$, 总平均值是

$$\hat{\mu} = \bar{X} = \frac{10\ 000 + 13\ 000 + 18\ 000 + 21\ 000 + 17\ 000}{425} = 185.88.$$

现在还不能计算 (16.64) 中 μ 的另一个估计, 稍后将给出答案.

经计算有

$$\begin{aligned}\hat{v} &= \frac{\begin{array}{c}50(200 - 209.09)^2 + 60(216.67 - 209.09)^2 + 100(180 - 177.78)^2 \\ +110(190.91 - 177.78)^2 + 105(161.90 - 177.78)^2\end{array}}{(2-1)+(3-1)} \\ &= 17\ 837.87\end{aligned}$$

以及

$$\begin{aligned}\hat{a} &= \frac{110(209.09 - 185.88)^2 + 315(177.78 - 158.88)^2 - (17\ 837.87)(1)}{425 - (110^2 + 315^2)/425} \\ &= 380.76,\end{aligned}$$

故 $\hat{k} = \hat{v}/\hat{a} = 46.85$. 信度因子的估计分别是

$$\hat{Z}_1 = \frac{110}{110 + 46.85} = 0.70, \quad \hat{Z}_2 = \frac{315}{315 + 46.85} = 0.87.$$

投保人 1 中每个个体的信度保费的估计是

$$\hat{Z}_1\bar{X}_1 + (1 - \hat{Z}_1)\hat{\mu} = (0.70)(209.09) + (0.30)(185.88) = 202.13.$$

因此对整个团体的信度保费的估计是

$$75(202.13) = 15\ 159.75.$$

对投保人 2, 有

$$\hat{Z}_2\bar{X}_2 + (1 - \hat{Z}_2)\hat{\mu} = (0.87)(177.78) + (0.13)(185.88) = 178.83.$$

总信度保费的估计是

$$90(178.83) = 16\ 094.70.$$

使用信度加权平均法有

$$\hat{\mu} = \frac{0.70(209.09) + 0.87(177.78)}{0.70 + 0.87} = 191.74.$$

于是信度保费分别是

$$0.70(209.09) + 0.30(191.74) = 203.89, \quad 0.87(177.78) + 0.13(191.74) = 179.59.$$

过去总信度保费是 $110(203.89) + 315(179.59) = 78\,998.75$. 除去四舍五入引起的误差, 结果与实际总损失 79 000 相吻合. □

在以上分析中假定参数 μ, v 和 a 未知, 需要进行估计, 实际中未必总是如此. 另外, 前面假设了 $n_i > 1$ 和 $r > 1$. 如果 $n_i = 1$, 那么关于投保人 i 只有一个经验数据, 要求解条件方差 $v(\Theta_i)$ 乃至 v 的值就会变得困难. 类似地, 如果 $r = 1$, 也就是说只有一个投保人, 那么要获得关于条件期望的方差 a 的信息会比较困难. 在这些情形下需要利用更强的假设, 比如说, 关于一个或多个参数的信息 (e.g. 已知净保费或手册费率 μ), 或者是参数之间的相关条件, 当中隐含了参数之间的某种函数关系. (将在 16.5.2 节和 16.5.3 节中讨论)

为了阐明以上观点, 比方说假设手册费率 μ 已知, 还需要估计 a 和 v. 这时 (16.61) 仍然可作为 v 的无偏估计, 不论 μ 是否已知. (为什么这里 $\left[\sum\limits_{j=1}^{n_i} m_{ij}(X_{ij}-\mu)^2\right]/n_i$ 不是 v 的无偏估计?) 同理, (16.63) 仍然是 a 的无偏估计量. 然而, 如果已知 μ, 另一个 a 的无偏估计是

$$\tilde{a} = \sum_{i=1}^{r} \frac{m_i}{m}(\bar{X}_i - \mu)^2 - \frac{r}{m}\hat{v},$$

其中 $\hat{v}$ 由 (16.61) 给出. 无偏性的证明如下

$$\begin{aligned}
\mathrm{E}(\tilde{a}) &= \sum_{i=1}^{r} \frac{m_i}{m}\mathrm{E}[(\bar{X}_i - \mu)^2] - \frac{r}{m}\mathrm{E}(\hat{v}) \\
&= \sum_{i=1}^{r} \frac{m_i}{m}\mathrm{Var}(\bar{X}_i) - \frac{r}{m}v \\
&= \sum_{i=1}^{r} \frac{m_i}{m}\left(a + \frac{v}{m_i}\right) - \frac{r}{m}v = a.
\end{aligned}$$

如果只有一个投保人的数据, 则这一类的方法是必不可少的. (16.60) 给出了只根据投保人 i 的数据做出的 v 的估计和 a 的无偏估计

$$\tilde{a}_i = (\bar{X}_i - \mu)^2 - \frac{\hat{v}_i}{m_i} = (\bar{X}_i - \mu)^2 - \frac{\sum_{j=1}^{n_i} m_{ij}(X_{ij} - \bar{X}_i)^2}{m_i(n_i - 1)},$$

上式的无偏性是因为 $\mathrm{E}[(\bar{X}_i - \mu)^2] = \mathrm{Var}(\bar{X}_i) = a + v/m_i$ 以及 $\mathrm{E}(\hat{v}_i) = v$.

例 16.37 对一组投保人有如下数据:

	第 1 年	第 2 年	第 3 年
总索赔数	60 000	70 000	–
组中人数	125	150	200

如果每人每年的手册费率是 500, 估计第 3 年的总信度保费.

解 由题设条件有 (不妨设这一组是投保人 i)$m_{i1}=125$, $X_{i1}=60\ 000/125=480$, $m_{i2}=150$, $X_{i2}=70\ 000/150=466.67$, $m_i=m_{i1}+m_{i2}=275$, $\bar{X}_i=(60\ 000+70\ 000)/275=472.73$, 进而有

$$\hat{v}_i=\frac{125(480-472.73)^2+150(466.67-472.73)^2}{2-1}=12\ 115.15,$$

再由 $\mu=500$ 知 $\tilde{a}_i=(472.73-500)^2-(12\ 115.15/275)=699.60$. 于是 k 的估计是 $\hat{v}_i/\tilde{a}_i=17.32$. 信度因子的估计是 $m_i/(m_i+\hat{v}_i/\tilde{a}_i)=275/(275+17.32)=0.94$, 个人信度保费的估计是 $0.94(472.73)+0.06(500)=474.37$, 第 3 年总信度保费的估计是 $200(474.37)=94\ 874$. □

除非没有别的办法, 一般情况下不推荐只根据单个投保人的数据估计 a 和 v, 因为这时估计量 $\hat{v}_i$ 和 $\tilde{a}_i$ 极不稳定. 尤其是当只从单个观察值 $(\bar{X}_i)$ 去估计 a 时, 强烈建议估计之前设法获取更多的经验数据.

16.5.2 半参数估计

有时候会假设条件分布 $f_{X_{ij}|\Theta_i}(x_{ij}|\theta_i)$ 具有某种参数形式, 并且能够从手上掌握的信息或是从先验分布中确认该假设的合理性. 例如, 在处理索赔次数数据时, 可以假设第 i 个投保人第 j 年的索赔次数为 $m_{ij}X_{ij}$, 在给定 $\Theta_i=\theta_i$ 的条件下服从均值为 $m_{ij}\theta_i$ 的 Poisson 分布. 这时有 $\mathrm{E}(m_{ij}X_{ij}|\Theta_i)=\mathrm{Var}(m_{ij}X_{ij}|\Theta_i)=m_{ij}\Theta_i$, 意味着 $\mu(\Theta_i)=v(\Theta_i)=\Theta_i$, 进而有 $\mu=v$. 于是可以使用 $\hat{\mu}=\bar{X}$ 去估计 v, 而不再用 (16.61).

例 16.38 某年机动车险索赔次数的分布如下:

索赔数	被保险人数
0	1 563
1	271
2	32
3	7
4	2
总计	1 875

基于这个数据计算每个投保人下一年索赔次数的信度估计. 假设索赔次数服从条件 Poisson 分布.

解 假设现在有 $r = 1\,875$ 个投保人, 每人有 $n_i = 1$ 年的经验数据, 风险量 $m_{ij} = 1$. 对第 i 个投保人 $(i = 1, \cdots, 1\,875)$ 假设 $X_{i1}|\Theta_i = \theta_i$ 服从均值为 θ_i 的 Poisson 分布, 所以有 $\mu(\theta_i) = v(\theta_i) = \theta_i$ 和 $\mu = v$. 仿照例 16.34, 有

$$\bar{X} = \frac{1}{1\,875}\left(\sum_{i=1}^{1\,875} X_{i1}\right) = \frac{0(1\,563) + 1(271) + 2(32) + 3(7) + 4(2)}{1\,875} = 0.194.$$

现在有

$$\begin{aligned}\mathrm{Var}(X_{i1}) &= \mathrm{Var}[\mathrm{E}(X_{i1}|\Theta_i)] + \mathrm{E}[\mathrm{Var}(X_{i1}|\Theta_i)] \\ &= \mathrm{Var}[\mu(\Theta_i)] + \mathrm{E}[v(\Theta_i)] = a + v = a + \mu.\end{aligned}$$

因此 $a + v$ 的一个无偏估计是样本方差

$$\frac{\sum_{i=1}^{1\,875}(X_{i1} - \bar{X})^2}{1\,874} = \frac{\begin{array}{c}1\,563(0 - 0.194)^2 + 271(1 - 0.194)^2 \\ +32(2 - 0.194)^2 + 7(3 - 0.194)^2 + 2(4 - 0.194)^2\end{array}}{1\,874} = 0.226.$$

故有 $\hat{a} = 0.226 - 0.194 = 0.032$ 和 $\hat{k} = 0.194/0.032 = 6.06$, 信度因子 Z 是 $1/(1 + 6.06) = 0.14$. 于是个体投保人索赔次数的信度估计是 $(0.14)X_{i1} + (0.86)(0.194)$, 其中 X_{i1} 根据不同投保人取值为 0, 1, 2, 3, 4. □

注意到这个例子中 $\mu = v$, 所以只要有个体投保人一年的经验数据即可进行估计.

例 16.39 要考察团体中的个体发生索赔的概率 (例如, 团体人身保险). 对不同的团体这个概率不尽相同. 用 $m_{ij}X_{ij}$ 表示团体 i 在第 j 年的 m_{ij} 个人中, 曾经发生过索赔的人数. 试根据以上条件建立信度模型.

解 如果团体 i 的成员发生索赔的概率是 θ_i, 那么不妨假设 $m_{ij}X_{ij}$ 在给定 $\Theta_i = \theta_i$ 的条件下服从参数为 m_{ij} 和 θ_i 的二项分布. 于是有

$$\mathrm{E}(m_{ij}X_{ij}|\Theta_i) = m_{ij}\Theta_i, \quad \mathrm{Var}(m_{ij}X_{ij}|\Theta_i) = m_{ij}\Theta_i(1 - \Theta_i),$$

以及 $\mu(\Theta_i) = \Theta_i$, $v(\Theta_i) = \Theta_i(1 - \Theta_i)$. 因此

$$\begin{aligned}\mu = \mathrm{E}(\Theta_i), \quad v = \mu - \mathrm{E}[(\Theta_i)^2], \\ a = \mathrm{Var}(\Theta_i) = \mathrm{E}[(\Theta_i)^2] - \mu^2 = \mu - v - \mu^2.\end{aligned}$$ □

在上述例子中作出了参数形式的假设, 使得在参数 μ, v 和 a 之间存在某种函数关系, 这经常会让它们的估计变得容易.

16.5.3 参数估计

如果对 $f_{X_{ij}|\Theta}(x_{ij}|\theta_i)$ 和 $\pi(\theta_i)$, $i=1,\cdots,r$, $j=1,\cdots,n_i$ 均作出参数形式的假设, 那么除了非参数估计的方法外, 还可以运用一整套的参数估计方法. 其中, 最大似然估计是一种比较直接的方法 (至少在理论上), 下面将进行讨论. 对投保人 i, $\boldsymbol{X}_i=(X_{i1},\cdots,X_{in_i})^{\mathrm{T}}(i=1,\cdots,r)$ 在给定 $\Theta_i=\theta_i$ 的条件下的联合密度是

$$f_{\boldsymbol{X}_i}(\boldsymbol{x}_i)=\int\left[\prod_{j=1}^{n_i}f_{X_{ij}|\Theta}(x_{ij}|\theta_i)\right]\pi(\theta_i)\mathrm{d}\theta_i. \tag{16.65}$$

似然函数是

$$L=\prod_{i=1}^{r}f_{\boldsymbol{X}_i}(\boldsymbol{X}_i). \tag{16.66}$$

最大似然估计就是使 L 或者 $\ln L$ 取到最大值时的参数值.

例 16.40 这是一个简单的例子. 假设对 $i=1,\cdots,r$ 有 $n_i=n$ 和 $m_{ij}=1$, 并且 $X_{ij}|\Theta_i$ 服从均值是 Θ_i 的 Poisson 分布, 也就是

$$f_{X_{ij}|\Theta}(x_{ij}|\theta_i)=\frac{\theta_i^{x_{ij}}\mathrm{e}^{-\theta_i}}{x_{ij}!},\quad x_{ij}=0,1,\cdots,$$

同时令 Θ_i 服从均值为 μ 的指数分布,

$$\pi(\theta_i)=\frac{1}{\mu}\mathrm{e}^{-\theta_i/\mu},\quad \theta_i>0.$$

求 μ 的最大似然估计.

解 等式 (16.65) 变成

$$\begin{aligned}f_{\boldsymbol{X}_i}(\boldsymbol{x}_i)&=\int_0^{\infty}\left(\prod_{j=1}^{n}\frac{\theta_i^{x_{ij}}\mathrm{e}^{-\theta_i}}{x_{ij}!}\right)\frac{1}{\mu}\mathrm{e}^{-\theta_i/\mu}\mathrm{d}\theta_i\\&=\left(\prod_{j=1}^{n}x_{ij}!\right)^{-1}\frac{1}{\mu}\int_0^{\infty}\theta_i^{\sum\limits_{j=1}^{n}x_{ij}}\mathrm{e}^{-\theta_i(n+1/\mu)}\mathrm{d}\theta_i\\&=C(\boldsymbol{x}_i)\mu^{-1}\left(n+\frac{1}{\mu}\right)^{-\sum\limits_{j=1}^{n}x_{ij}-1}\int_0^{\infty}\frac{\beta(\beta\theta_i)^{\alpha-1}\mathrm{e}^{-\beta\theta_i}}{\Gamma(\alpha)}\mathrm{d}\theta_i,\end{aligned}$$

其中 $C(\mathbf{x}_i)$ 可以用组合记号表示成

$$C(\boldsymbol{x}_i)=\binom{\sum\limits_{j=1}^{n}x_{ij}}{x_{i1}x_{i2}\cdots x_{ij}};\quad \beta=n+\frac{1}{\mu},$$

且有

$$\alpha = \sum_{j=1}^{n} x_{ij} + 1.$$

积分号内是参数为 α 和 $1/\beta$ 的 gamma 分布的密度函数, 因此积分值为 1, 故有

$$f(\boldsymbol{x}_i) = C(\boldsymbol{x}_i)\mu^{-1}\left(n + \frac{1}{\mu}\right)^{-\sum\limits_{j=1}^{n} x_{ij} - 1}.$$

代入 (16.66) 得到

$$L(\mu) \propto \mu^{-r}\left(n + \frac{1}{\mu}\right)^{-\sum\limits_{i=1}^{r}\sum\limits_{j=1}^{n} x_{ij} - r}.$$

因此有

$$l(\mu) = \ln L(\mu) = -r\ln\mu - \left(r + \sum_{i=1}^{r}\sum_{j=1}^{n} x_{ij}\right)\ln\left(n + \frac{1}{\mu}\right) + c,$$

其中 c 是不依赖 μ 的常数. 对上式求导得到

$$l'(\mu) = -\frac{r}{\mu} - \frac{r + \sum_{i=1}^{r}\sum_{j=1}^{n} x_{ij}}{n + \dfrac{1}{\mu}}\left(-\frac{1}{\mu^2}\right).$$

μ 的最大似然估计 $\hat{\mu}$ 满足 $l'(\hat{\mu}) = 0$, 也即

$$\frac{r}{\hat{\mu}} = \frac{r + \sum_{i=1}^{r}\sum_{j=1}^{n} x_{ij}}{\hat{\mu}(\hat{\mu}n + 1)},$$

故有

$$\hat{\mu}n + 1 = 1 + \frac{1}{r}\sum_{i=1}^{r}\sum_{j=1}^{n} x_{ij},$$

从而

$$\hat{\mu} = \frac{1}{nr}\sum_{i=1}^{r}\sum_{j=1}^{n} x_{ij}.$$

这也就是例 16.34 中非参数估计的结果. 下面给出一个解释, 由 Poisson 分布的假设知 $\mu(\theta_i) = \theta_i$, 所以 $\mathrm{E}[\mu(\Theta_i)] = \mathrm{E}(\Theta_i)$, 与指数分布 $\pi(\theta_i)$ 中的 μ 相同.

此外, 由 Poisson 分布的假定知 $v(\theta_i) = \theta_i$, 所以 $v = \mathrm{E}[v(\Theta_i)] = \mu$. 并且利用 $\pi(\theta_i)$ 是指数分布的假设知 $a = \mathrm{Var}[\mu(\Theta_i)] = \mathrm{Var}(\Theta_i) = \mu^2$. 因此, 根据最大似然估计在参数变换下的不变性, v 和 a 的最大似然估计分别是 $\hat{\mu}$ 和 $\hat{\mu}^2$. 同理, $k = v/a$, 信度因子 Z, 还有信度保费 $Z\bar{X}_i + (1-Z)\mu$ 的最大似然估计分别是 $\hat{k} = \hat{\mu}^{-1} = \bar{X}^{-1}$,

$\hat{Z} = n/(n + \hat{\mu}^{-1})$ 以及 $\hat{Z}\bar{X}_i + (1 - \hat{Z})\hat{\mu}$. 注意到在这个模型中满足精确信度, 所以贝叶斯保费等于信度保费. □

例 16.41 假设对所有 i 有 $n_i = n$, $m_{ij} = 1$, 还假定 $X_{ij}|\Theta_i \sim N(\Theta_i, v)$,

$$f_{X_{ij}|\Theta}(x_{ij}|\theta_i) = (2\pi v)^{-1/2} \exp\left[-\frac{1}{2v}(x_{ij} - \theta_i)^2\right], \quad -\infty < x_{ij} < \infty,$$

以及 $\Theta_i \sim N(\mu, a)$, 即

$$\pi(\theta_i) = (2\pi a)^{-1/2} \exp\left[-\frac{1}{2a}(\theta_i - \mu)^2\right], \quad -\infty < \theta_i < \infty.$$

确定诸参数的最大似然估计.

解 由 $\mu(\theta_i) = \theta_i$ 和 $v(\theta_i) = v$, 有 $\mu = \mathrm{E}[\mu(\Theta_i)], v = \mathrm{E}[v(\Theta_i)]$ 和 $a = \mathrm{Var}[\mu(\Theta_i)]$, 和以前使用的 μ, v, a 保持一致. 现在来推导 μ, v 和 a 的最大似然估计. 考虑 $\bar{X}_i = n^{-1}\sum\limits_{j=1}^{n} X_{ij}$, 在给定 Θ_i 的条件下 X_{ij} 是相互独立的 $N(\Theta_i, v)$ 随机变量, 这意味着 $\bar{X}_i|\Theta_i \sim N(\Theta_i, v/n)$. 因为 $\Theta_i \sim N(\mu, a)$, 由例 4.30 可知在无条件情形下有 $\bar{X}_i \sim N(\mu, a + v/n)$. 所以 $\bar{X}_i$ 的密度是 (令 $w = a + v/n$)

$$f(\bar{x}_i) = (2\pi w)^{-1/2} \exp\left[-\frac{1}{2w}(\bar{x}_i - \mu)^2\right], \quad -\infty < \bar{x}_i < \infty.$$

另一方面, 利用给定 Θ_i 的条件密度, 有

$$f(\bar{x}_i) = \int_{-\infty}^{\infty} (2\pi v/n)^{-1/2} \exp\left[-\frac{n}{2v}(\bar{x}_i - \theta_i)^2\right] \times (2\pi a)^{-1/2} \exp\left[-\frac{1}{2a}(\theta_i - \mu)^2\right] \mathrm{d}\theta_i.$$

忽略那些不含 μ, v 和 a 的项, 这意味着 $f(\bar{x}_i)$ 和下面的项成比例:

$$v^{-1/2}a^{-1/2}\int_{-\infty}^{\infty} \exp\left[-\frac{n}{2v}(\bar{x}_i - \theta_i)^2 - \frac{1}{2a}(\theta_i - \mu)^2\right] \mathrm{d}\theta_i.$$

利用 (16.65) 得到

$$f(\boldsymbol{x}_i) = \int_{-\infty}^{\infty} \left\{\prod_{j=1}^{n} (2\pi v)^{-1/2} \exp\left[-\frac{1}{2v}(x_{ij} - \theta_i)^2\right]\right\} (2\pi a)^{-1/2} \times \exp\left[-\frac{1}{2a}(\theta_i - \mu)^2\right] \mathrm{d}\theta_i,$$

它与

$$v^{-n/2}a^{-1/2}\int_{-\infty}^{\infty} \exp\left[-\frac{1}{2v}\sum_{j=1}^{n}(x_{ij} - \theta_i)^2 - \frac{1}{2a}(\theta_i - \mu)^2\right] \mathrm{d}\theta_i.$$

成比例. 由 (16.10) 可得

$$\sum_{j=1}^{n}(x_{ij}-\theta_i)^2=\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2+n(\bar{x}_i-\theta_i)^2,$$

也就是说 $f(\mathbf{x}_i)$ 与

$$v^{-n/2}a^{-1/2}\int_{-\infty}^{\infty}\exp\left\{-\frac{1}{2v}\left[\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2+n(\bar{x}_i-\theta_i)^2\right]-\frac{1}{2a}(\theta_i-\mu)^2\right\}\mathrm{d}\theta_i,$$

或是与

$$v^{-(n-1)/2}\exp\left[-\frac{1}{2v}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2\right]f(\bar{x}_i)$$

成比例, 当中用到了 $f(\bar{x}_i)$ 的第二个表达式. 因此由 (16.66) 可以得到

$$L\propto v^{-r(n-1)/2}\exp\left[-\frac{1}{2v}\sum_{i=1}^{r}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2\right]\prod_{i=1}^{r}f(\bar{x}_i).$$

借助于参数变换下最大似然估计的不变性, 用参数 μ, v 和 $w=a+v/n$, 而不是 μ, v 和 a 来讨论. 上式等价于

$$L\propto L_1(v)L_2(\mu,w),$$

其中

$$L_1(v)=v^{-r(n-1)/2}\exp\left[-\frac{1}{2v}\sum_{i=1}^{r}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2\right],$$

$$L_2(\mu,w)=\prod_{i=1}^{r}f(\bar{x}_i)=\prod_{i=1}^{r}\left\{(2\pi w)^{-1/2}\exp\left[-\frac{1}{2w}(\bar{x}_i-\mu)^2\right]\right\}.$$

v 的最大似然估计 $\hat{v}$ 可以通过最大化 $L_1(v)$ 得到, 并且 (μ, w) 的最大似然估计 $(\hat{\mu},\hat{w})$ 可以通过最大化 $L_2(\mu,w)$ 得到. 取对数后, 有

$$l_1(v)=-\frac{r(n-1)}{2}\ln v-\frac{1}{2v}\sum_{i=1}^{r}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2,$$

$$l_1'(v)=-\frac{r(n-1)}{2v}+\frac{1}{2v^2}\sum_{i=1}^{r}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2,$$

令 $l'(\hat{v})=0$ 可得

$$\hat{v}=\frac{\sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij}-\bar{X}_i)^2}{r(n-1)}.$$

由于 $L_2(\mu, w)$ 是普通的正态似然函数, 因此最大似然估计就是经验均值和方差, 也就是

$$\hat{\mu} = \frac{1}{r}\sum_{i=1}^{r}\bar{X}_i = \frac{1}{nr}\sum_{i=1}^{r}\sum_{j=1}^{n}X_{ij} = \bar{X}, \quad \hat{w} = \frac{1}{r}\sum_{i=1}^{r}(\bar{X}_i - \bar{X})^2.$$

再由 $a = w - v/n$ 可知 a 的最大似然估计是

$$\hat{a} = \frac{1}{r}\sum_{i=1}^{r}(\bar{X}_i - \bar{X})^2 - \frac{1}{rn(n-1)}\sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2.$$

值得注意的是, 最大似然估计 $\hat{\mu}$ 和 $\hat{v}$ 正好等于例 (16.34) 中 Bühlmann 模型的非参数无偏估计. 最大似然估计 $\hat{a}$ 和它的非参数无偏估计几乎相同, 除了在第一项中把 $r-1$ 换成了 r. □

16.5.4 备注

本节采用了一种简单方法进行参数估计, 而没有尝试寻找最小方差意义下的最佳估计. 对这个问题进行了许多相关的研究, 可参见 Goovaerts and Hoogstad[46] 得到更详细的论述和进一步的参考.

习题

16.60 表 16-8 给出了一组投保人的过去索赔数据. 估计每个投保人第 4 年的 Bühlmann 信度保费.

表 16-8 习题 16.60 的数据

投保人团体	年份		
	1	2	3
1	750	800	650
2	625	600	675
3	900	950	850

16.61 表 16-9 给出了若干个团体过去的索赔数据. 估计第 4 年应向每个团体收取的 Bühlmann-Straub 信度保费.

表 16-9 习题 16.61 的数据

	投保人团体	年份			
		1	2	3	4
索赔数	1	–	20 000	25 000	–
团体中人数		–	100	120	110
索赔数	2	19 000	18 000	17 000	–
团体中人数		90	75	70	60
索赔数	3	26 000	30 000	35 000	–
团体中人数		150	175	180	200

16.62 题设条件同习题 16.9. 估计投保人下一年的 Bühlmann 信度保费.

16.63 考虑例 16.34 中的 Bühlmann 模型.

(a) 证明 $\mathrm{Var}(X_{ij}) = a + v$.

(b) 如果 $\{X_{ij} : i = 1, \cdots, r, j = 1, \cdots, n\}$ 对所有 i 和 j 无条件独立, 证明 $a + v$ 的一个无偏估计是

$$\frac{1}{nr-1}\sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij} - \bar{X})^2.$$

(c) 证明如下等式

$$\sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij} - \bar{X})^2 = \sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2 + n\sum_{i=1}^{r}(\bar{X}_i - \bar{X})^2.$$

(d) 证明在原来的题设条件下, 有

$$\mathrm{E}\left[\frac{1}{nr-1}\sum_{i=1}^{r}\sum_{j=1}^{n}(X_{ij} - \bar{X})^2\right] = (v + a) - \frac{n-1}{nr-1}a.$$

(e) 说明 (b) 和 (d) 的含义.

16.64 表 16-10 给出了汽车保险投保人的索赔次数分布. 假设每个投保人的索赔次数服从条件 Poisson 分布, 给出下一年索赔次数的 Bühlmann 信度保费.

表 16-10 习题 16.64 的数据

索赔数	投保人数
0	2 500
1	250
2	30
3	5
4	2
总计	2 787

16.65 假设给定 Θ 的条件下, $X_1, \cdots, X_n$ 是相互独立的几何分布, 概率函数是

$$f_{X_j|\Theta}(x_j|\theta) = \frac{1}{1+\theta}\left(\frac{\theta}{1+\theta}\right)^{x_j}, \quad x_j = 0, 1, \cdots.$$

(a) 证明 $\mu(\theta) = \theta, v(\theta) = \theta(1+\theta)$.

(b) 证明 $a = v - \mu - \mu^2$.

(c) 重新计算习题 16.64, 假定每个人索赔次数服从条件几何分布.

16.66 假设

$$\Pr(m_{ij}X_{ij} = t_{ij}|\Theta_i = \theta_i) = \frac{(m_{ij}\theta_i)^{t_{ij}}\mathrm{e}^{-m_{ij}\theta_i}}{t_{ij}!},$$

$$\pi(\theta_i) = \frac{1}{\mu}\mathrm{e}^{-\theta_i/\mu}, \quad \theta_i > 0.$$

对 Bühlmann-Straub 型数据, 写出 μ 的最大似然估计 $\hat{\mu}$ 满足的方程.

16.67 (a) 证明等式

$$\sum_{i=1}^{r}\sum_{j=1}^{n_i} m_{ij}(X_{ij}-\bar{X})^2=\sum_{i=1}^{r}\sum_{j=1}^{n_i} m_{ij}(X_{ij}-\bar{X}_i)^2+\sum_{i=1}^{r} m_i(\bar{X}_i-\bar{X})^2.$$

(b) 利用 (a) 和 (16.61) 证明 (16.63) 可以写为

$$\hat{a}=m_*^{-1}\left[\frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i} m_{ij}(X_{ij}-\bar{X})^2}{\sum_{i=1}^{r} n_i-1}-\hat{v}\right],$$

其中

$$m_*=\frac{\sum_{i=1}^{r} m_i\left(1-\frac{m_i}{m}\right)}{\sum_{i=1}^{r} n_i-1}.$$

16.68* 表 16-11 给出了一年中来自罪案高发地区的 340 个投保人所上报的 210 次盗窃案.

表 16-11 习题 16.68 的数据

索赔数	被保险人数
0	200
1	80
2	50
3	10

假设每个被保险人的被盗次数服从 Poisson 分布, 且不同被保险人该分布的均值不尽相同. 如果某被保险人在观察期内遭受了两次盗窃, 求此被保险人下一年被盗次数的 Bühlmann 信度估计.

第17章　随机模拟

17.1　随机模拟的基础知识

在精算实践的历史上, 随机模拟 (simulation) 曾一度被采用, 又一度被抛弃. 举例来说, 在 20 世纪 70 年代, 总体损失的计算通常是采用模拟方法, 因为解析方法并不令人满意. 然而, 即使简单的模拟通常也需要公司主机运行一整天的时间, 造成了严重的资源占用. 到了 20 世纪 80 年代, 人们开发了诸如 Heckman-Meyers 方法以及递推公式等解析方法. 这些方法被证实比计算机模拟快得多而且更加精确. 现在的台式计算机也已经足够强大可以进行复杂的模拟, 这使得一些目前解析方法不适用的模型分析得以实现.

在类似的思路下, 当投资工具变得更加复杂、合同具有利率敏感性, 以及市场波动变得更加显著时, 对未来现金流的分析必须建立在随机基础之上. 为了适应产品的复杂性和利率模型, 随机模拟已成为人们必选的工具.

本章将说明如何利用随机模拟解决上面提及的问题. 目的不在于详细讨论模拟的技术细节, 而是要给读者一些关于模拟是如何帮助解决问题的概念. 其他关于模拟的教材, 例如 Herzog and Lord [53], Ripley[110], 以及 Ross[115]①会提供更多相关的重要细节. 另外, 模拟也可以帮助我们评价前面章节中提到的一些统计方法, 这一点也将在本章以自助法为重点进行讨论.

模拟的方法

模拟方法的优点是一旦模型建立起来, 就不需要什么其他创造性的想法②. 当模拟的目标是给出某个随机变量 S 的分布时, 整个模拟过程可以总结为下面 4 个步骤.

(1) 建立 S 的模型. 这个模型依赖于随机变量 $X,Y,Z,\cdots$, 而这些随机变量的分布以及它们之间的依赖关系是已知的.

(2) 对 $j=1,\cdots,n$ 生成伪随机数 $x_j,y_j,z_j,\cdots$, 然后由步骤 (1) 中的模型计算 s_j.

---

① 中文版和英文影印版《随机模拟》已由人民邮电出版社出版. —— 编者注

② 这一点并不完全正确. 在设计一个有效的模拟算法时需要许多的创造性工作. 采用强力方法 (brute force approach) 也可以解决问题; 只是需要耗费计算机更多的时间才能得到满意的结果.

(3) 用 $F_n(s)$ 近似 S 的累积分布函数, 其中 $F_n(s)$ 是伪随机样本 $s_1, \cdots, s_n$ 的经验分布函数.

(4) 利用经验分布函数计算其他的量, 比如均值、方差、分位点或者某些概率值.

这里有两个问题. 第一个问题是: 生成的伪随机变量到底意味着什么? 考虑某随机变量 X, 累积分布函数为 $F_X(x)$. 它代表所关心的某种现象生成的真实随机变量, 比如, 它可以是"随机收集的机动车人身伤害医疗赔付记录". 假设累积分布函数是已知的, 例如 Pareto 分布, $F_X(x) = 1 - (\frac{1\,000}{1\,000+x})^3$. 现在考虑另外一个随机变量 X^*, 生成自某个其他的过程, 但是服从相同的 Pareto 分布. 无法区分来自 X^* 的随机样本 $x_1^*, \cdots, x_n^*$, 和另一个来自 X 的样本. 也就是说, 给定 n 个随机数, 不知道它们来自机动车险的索赔还是来自某个其他的过程. 这意味着, 可以通过观测 X^* 了解我们关心的现象, 而不必观测机动车险的索赔. 获得 Pareto 分布的随机样本仍然很困难, 因此目前还没有很大进展.

通过做出一些让步, 可以取得一些进展. 接受用数列 $x_1^{**}, \cdots, x_n^{**}$ 代替 X^* 的随机样本, 虽然这个数列根本就不是随机序列, 而只是一些不独立甚至非随机的数构成的序列, 但是这个序列是通过与随机变量 X^* 相关的某个过程生成的. 称这样一个序列为伪随机序列, 因为任何不知其来源的人都不能将它与 X^* 的随机样本 (因此还有 X 的随机样本) 相区分. 这样一个序列已经能够满足我们的要求.

关于伪随机序列的生成方法已经发展得很完善了, 为了让随机数的生成变得很简单, 只要生成 (0,1) 区间均匀分布的随机数就足够了. 这是因为, 如果 U 服从 (0,1) 区间的均匀分布, 则 $X = F_X^{-1}(U)$(如果 $F_X(x)$ 的逆函数有明确的定义) 的累积分布函数就是 $F_X(x)$. 因此可以由均匀分布的伪随机数 $u_1^{**}, \cdots, u_n^{**}$ 得到 $x_j^{**} = F_X^{-1}(u_j^{**})$, 称之为生成随机变量的逆变换法. 已经得到了一些针对特殊分布的具体方法, 但这里将不作讨论. 有很多关于生成均匀分布伪随机数的最好方法以及评价它们的各种检验的文献. 读者应该仔细确认所使用的方法的优良性.

例 17.1 *生成 10 000 个伪 Pareto 随机数 (参数 $\alpha = 3, \theta = 1\,000$) 并且验证无法将其与真实的 Pareto 分布的观测值区分开.*

解 由某个商业化的程序语言内置的生成器得到均匀分布的伪随机数. 然后由下面公式得到伪 Pareto 随机数

$$u^{**} = 1 - \left(\frac{1\,000}{1\,000 + x^{**}}\right)^3.$$

即

$$x^{**} = 1\,000[(1 - u^{**})^{-1/3} - 1].$$

因此, 如果第一个生成值为 $u_1^{**} = 0.542\,46$, 则有 $x_1^{**} = 297.75$. 重复这个步骤

10 000 次, 结果汇总于表 17-1, 并给出了 χ^2 拟合优度检验的结果. 表中期望数一列是用参数 $\alpha = 3, \theta = 1\ 000$ 的 Pareto 分布计算得到的, 因为参数是已知的, 所以有 9 个自由度. 5%显著水平的临界值为 16.92, 因此可以认为伪随机样本是 Pareto 分布的一个随机样本. □

表 17-1　模拟生成 Pareto 观测值的 χ^2 拟合优度检验

区　间	观测值	期望	卡方
0~100	2 519	2 486.85	0.42
100~250	2 348	2 393.15	0.85
250~500	2 196	2 157.04	0.70
500~750	1 071	1 097.07	0.62
750~1 000	635	615.89	0.59
1 000~1 500	589	610.00	0.72
1 500~2 500	409	406.76	0.01
2 500~5 000	192	186.94	0.14
5 000~10 000	36	38.78	0.20
10 000~	5	7.51	0.84
总计	10 000	10 000	5.10

当 X 的分布是连续且严格单调递增时, 方程 $u = F_X(x)$ 对于任何给定的 u 都只有唯一解. 这时逆变换法简化为解方程, 而其他情况下就必须小心了. 假设 $F_X(x)$ 在 $x = c$ 处有跳跃, 使得 $F_X(c-) = a$, $F_X(c) = b > a$. 当均匀分布的随机数满足 $a \leqslant u < b$ 时, 方程无解. 在这种情况下, 可选择 c 作为模拟值.

例 17.2　已知

$$F_X(x) = \begin{cases} 0.5x, & 0 \leqslant x < 1, \\ 0.5 + 0.25x & 1 \leqslant x \leqslant 2. \end{cases}$$

试给出由均匀分布随机数 0.3, 0.6, 0.9 得到的 x 模拟值.

解　在第一个区间中, 分布函数从 0 到 0.5 取值, 在第二个区间中从 0.75 到 1 取值. 对于处于第一个区间中的 $u = 0.3$, 解方程 $0.3 = 0.5x$ 得到 $x = 0.6$. 在 $x = 1$ 处分布函数从 0.5 跳跃至 0.75, 对这个区间中的每个 u, x 的模拟值都是 1. 因此对于 $u = 0.6$, 模拟值是 $x = 1$. 注意到 $\Pr(0.5 \leqslant U < 0.75) = 0.25$, 因此模拟结果中有 25%是 $x = 1$, 与它的真实概率相对应. 最后, 0.9 在第二个区间中, 解方程 $0.9 = 0.5 + 0.25x$ 得到 $x = 1.6$. 图 17-1 给出了这个过程并显示了如何在分布函数上作垂线使得逆变换更明显. □

分布函数在某个区间上取值为常数也是有可能的. 在这种情况下, 方程 $u = F_X(x)$ 在那个区间上会有多个解. 我们的惯例 (马上将会证明这是合理的) 是取那个区间中最大可能的值.

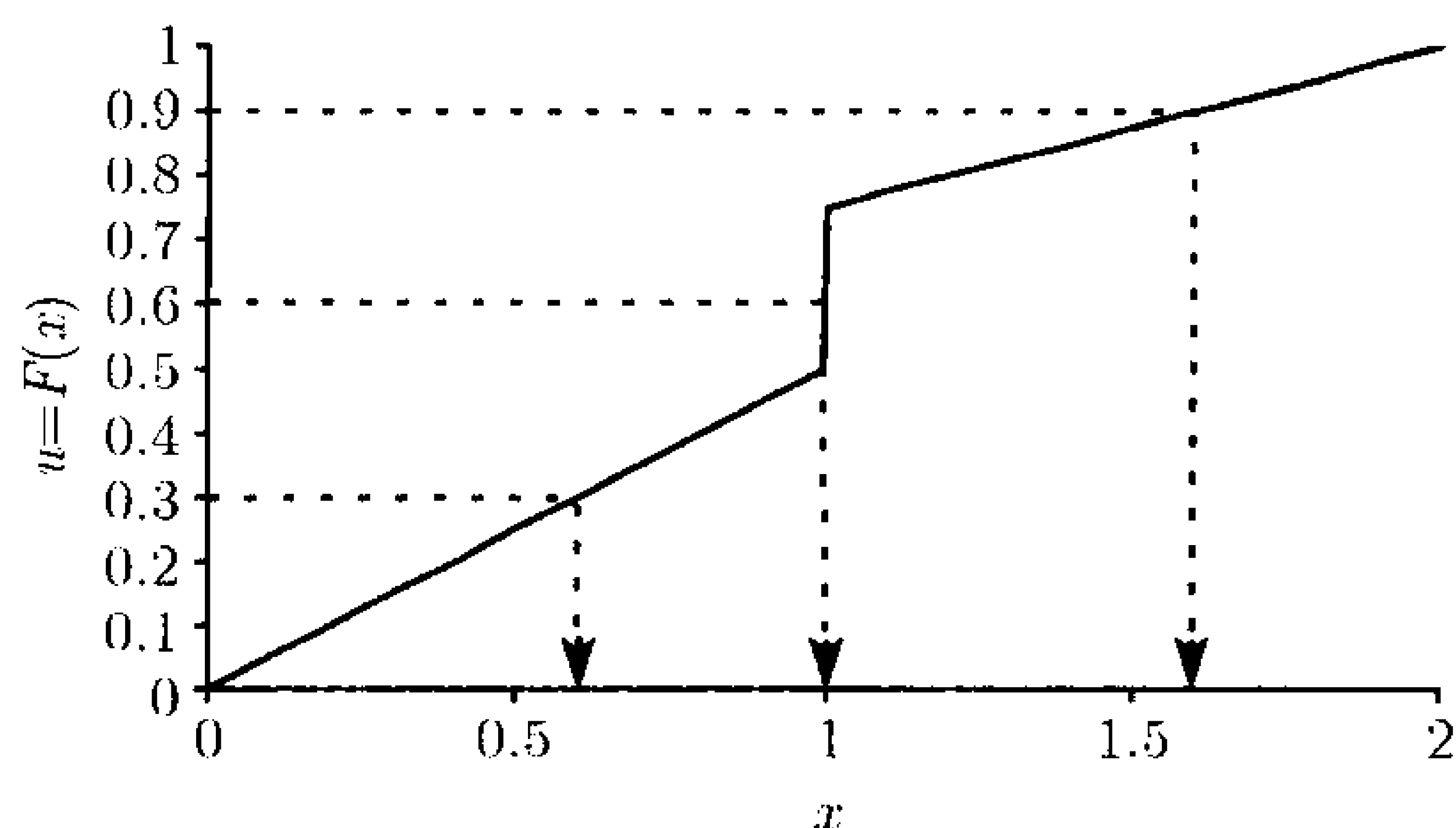


图 17-1 例 17.2 分布函数的逆变换

例 17.3 已知

$$F_X(x)=\begin{cases}0.5x, & 0\leqslant x<1,\\ 0.5, & 1\leqslant x<2,\\ 0.5x-0.5, & 2\leqslant x<3.\end{cases}$$

试给出由均匀分布随机数 0.3, 0.5, 0.9 得到的 x 的模拟值.

解 在第一个区间中, 分布函数取值从 0 到 0.5. 最后一个区间的范围是 0.5 到 1. 对于 $u=0.3$, 用第一个区间, 解方程 $0.3=0.5x$ 得到 $x=0.6$. 分布函数在 1 到 2 上取值恒为 0.5, 因此对 $u=0.5$, 选最大值 $x=2$ 作为模拟值. 对 $u=0.9$, 用最后一个区间, 解方程 $0.9=0.5x-0.5$ 得到 $x=2.8$. □

离散分布具有以下两种特征：分布函数在随机变量的所有可能取值处跳跃并在这些值之间取常数.

例 17.4 **用均匀分布随机数 0.3, 0.687 5, 0.95 生成参数为 $m=4, q=0.5$ 的二项分布的模拟值.**

解 分布函数为

$$F_X(x)=\begin{cases}0, & x<0,\\ 0.062\,5, & 0\leqslant x<1,\\ 0.312\,5, & 1\leqslant x<2,\\ 0.687\,5, & 2\leqslant x<3,\\ 0.937\,5, & 3\leqslant x<4,\\ 1, & x\geqslant 4.\end{cases}$$

当 $u=0.3$ 时, 函数在 $x=1$ 处跳跃. 对于 $u=0.687\,5$, 由于函数在 2 到 3(区间的端点) 上为常数, 因此 $x=3$. 对于 $u=0.95$, 函数在 $x=4$ 处跳跃. 分布函数表通常能够更好地说明这种模拟算法, 然后用查表函数 (比如 Excel 中的 VLOOKUP 函数) 得到模拟值. 对于这个例子, 分布函数表如下所示.

u 所在区间	模拟值
$0 \leqslant u < 0.062\ 5$,	0
$0.062\ 5 \leqslant u < 0.312\ 5$,	1
$0.312\ 5 \leqslant u < 0.687\ 5$,	2
$0.687\ 5 \leqslant u < 0.937\ 5$,	3
$0.937\ 5 \leqslant u < 1$,	4

许多随机数生成器都可以生成数字 0 但不能生成数字 1(尽管有些程序两者都无法生成). 这也是我们对分布函数为常数时选择最大值的原因. □

第二个问题是：如何选取 n? 当样本容量增大时, 相合估计将以很大的概率任意靠近真实值. 特别地, 经验估计量具有这种性质. 通过一些努力, 应该可以确定 n 的值, 使得对于某个给定的概率, 估计量能够达到要求的任意精度. 如下例所示, 中心极限定理经常能够提供这方面的帮助.

例 17.5(续例 17.1) *利用随机模拟估计参数为 $\alpha = 3, \theta = 1\ 000$ 的 Pareto 分布的均值、$F_X(1\ 000)$ 以及 90%分位点 $\pi_{0.9}$. 当你确信有 95%的概率得到的模拟结果与真实值的误差为 $\pm 1\%$ 时停止模拟.*

解 在此例中, 实际上已知 $\mu = 500$, $F_X(1\ 000) = 0.875$, $\pi_{0.9} = 1\ 154.43$, 但为了说明模拟方法, 假设并不知道这些值.

μ 的经验估计为 $\bar{x}$. 由中心极限定理, 在样本容量为 n 时, 有

$$\begin{aligned} 0.95 &= \Pr(0.99\mu \leqslant \bar{X}_n \leqslant 1.01\mu) \\ &= \Pr\left(-\frac{0.01\mu}{\sigma/\sqrt{n}} \leqslant \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leqslant \frac{0.01\mu}{\sigma/\sqrt{n}}\right) \\ &= \Pr\left(-\frac{0.01\mu}{\sigma/\sqrt{n}} \leqslant Z \leqslant \frac{0.01\mu}{\sigma/\sqrt{n}}\right), \end{aligned}$$

其中 Z 服从标准正态分布. 当下式成立时, 就可达到我们的目标

$$\frac{0.01\mu}{\sigma/\sqrt{n}} = 1.96, \tag{17.1}$$

这意味着 $n = 38\ 416(\sigma/\mu)^2$. 因为不知道 σ 和 μ 的真值, 故用样本标准差和样本均值进行估计. 这些估计量随着 n 增加而得到改进, 因此停止准则是当下式成立时停止模拟

$$n \geqslant \frac{38\ 416 s^2}{\bar{x}^2}.$$

在本书进行的一次模拟中, 当 $n = 106\ 934$ 时上述准则满足. 这时 $\bar{x} = 501.15$, 相对误差是 0.23%, 达到了我们的目标.

现在估计 $F_X(1\ 000)$. 经验估计量为样本中小于 1 000 的观测值的比例 P_n/n, 其中 P_n 是 n 次模拟中小于 1 000 的观测数. 由中心极限定理, P_n/n 渐进地服从均

值为 $F_X(1\ 000)$, 方差为 $F_X(1\ 000)[1-F_X(1\ 000)]/n$ 的正态分布. 类似于上面的讨论, 当

$$n \geqslant 38\ 416\frac{n-P_n}{P_n}.$$

成立时, 满足要求. 在模拟中, 当 $n=5\ 548$ 时满足标准, 这时的估计为 $4\ 848/5\ 548=0.873\ 83$, 相对误差是 0.13%.

最后是 $\pi_{0.9}$ 的估计, 初值为

$$0.95=\Pr(Y_a \leqslant \pi_{0.9} \leqslant Y_b),$$

其中 $Y_1 \leqslant Y_2 \leqslant \cdots \leqslant Y_n$ 是模拟样本的次序统计量, a 是小于等于 $0.9n+0.5-1.96\sqrt{0.9(0.1)n}$ 的最大整数, b 是大于等于 $0.9n+0.5+1.96\sqrt{0.9(0.1)n}$ 的最小整数. 当

$$\hat{\pi}_{0.9}-Y_a \leqslant 0.01\hat{\pi}_{0.9}, \quad Y_b-\hat{\pi}_{0.9} \leqslant 0.01\hat{\pi}_{0.9}.$$

同时成立时, 过程停止. 在本例中, 当 $n=126\ 364$ 时过程中止, 90%分位点的估计值为 1 153.97, 相对误差是 0.04%. □

习题

17.1 利用逆变换法生成三个 Poisson(3) 分布的模拟值. 用 0.124 7, 0.932 1, 0.687 3 作为均匀分布随机数.

17.2 利用均匀分布的随机数 0.2, 0.5, 0.7 生成服从以下分布的随机数

$$f_X(x)=\begin{cases} 0.25, & 0 \leqslant x \leqslant 2, \\ 0.1, & 4 \leqslant x \leqslant 9, \\ 0, & \text{其他}. \end{cases}$$

17.3 证明例 17.5 中定义的 Y_a 和 Y_b 满足 $0.95=\Pr(Y_a \leqslant \pi_{0.9} \leqslant Y_b)$.

17.4 假设由 $\theta=100$ 的指数分布生成随机数. 问: 需要多少次模拟才能够在 90%置信水平下, 均值以及小于 200 的概率的估计值在真实值 2%左右? 计算所需的次数并且验证 2%的目标是否得到满足.

17.5 从参数为 $\alpha=2, \theta=500$ 的 gamma 分布中生成 1 000 个观测. 利用 χ^2 拟合优度检验和 Komolgorov-Smirnov 检验来验证模拟值是否真正来自所给分布.

17.6* 为了估计 $\mathrm{E}(X)$, 模拟生成了随机变量 X 的 5 个观测, 值为 1,2,3,4,5. 若目标是使得 $\mathrm{E}(X)$ 估计量的标准差小于 0.05, 估计需要模拟的总次数.

17.2　精算建模中的随机模拟实例

17.2.1　总体损失计算

第 6 章介绍的解析方法有两个共同的特点. 首先, 严格依赖于近似的程度. 递推和 FFT 需要用数值估计来代替真实的赔额分布. 对于 Heckman-Meyers 方法, 需要直方图近似, 另外, 还需要计算数值积分. 在任何情况下, 都可以通过增加点的个数, 将误差减小到接近于 0. 其次, 递推和逆变换方法均假设总索赔额可以表示为 $S = X_1 + \cdots + X_N$, 其中 $N, X_1, \cdots, X_n$ 相互独立并且 X_j 是同分布的.

没有必要关心上面提到的第一个问题, 因为近似误差可以减小到我们要求的程度. 然而, 第二个局限性可能会使模型不能反映现实情况. 本节将指出独立或者同分布假定不能成立的常见情形, 然后展示模拟方法如何给出这些问题的解答. 当所有 X_j 是独立同分布时, 我们并不在意损失的标号, 即哪个是 X_1, 哪个是 X_2 并不重要. 但当独立性假定去掉后, 标号就变得重要了. 因为 S 表示一年的总损失, 损失发生时间是一个重要因素. 一种区分方法是令 X_1 表示第一个损失, X_2 表示第二个损失, 依此类推. 然后令 T_j 表示第 j 个损失发生的时间. 这里没有详细处理具体的赔付过程, 需要提醒读者注意的是 T_j 可能是损失发生的时间、报告的时间或者是赔付的时间. 在后两种情况下, 也可能有 $T_j > 1$, 即损失的报告或者赔付发生在承保进行的下一年. 如果损失发生时间是重要的, 就需要知道 $(T_1, T_2, \cdots, X_1, X_2, \cdots)$ 的联合分布.

17.2.2　无独立性或同分布假设的例子

独立性或同分布假设在以下两种常见的情形下不成立. 一是考虑时间 (特别地, 货币的时间价值), 另一个是承保责任的调整. 在后一种情况下可能也有时间因素. 下面的例子给出了具体说明.

例 17.6(损失赔付的时间价值)　*假设所关心的量 S 是今天发行的有效期为 1 年的保单的所有赔付的现值. 对 S 建模.*

解　用 T_j 表示第 j 笔赔付发生的时间, 下标按照损失发生的次序排序. $T_j = C_j + L_j$, 其中 C_j 是事件发生的时间, L_j 表示从损失发生到赔付结案时间. 假设 C_j 与 L_j 之间是独立的并且 L_j 之间相互独立. 设事件发生间隔 $C_j - C_{j-1} (C_0 = 0)$ 是独立同分布的随机变量, 服从期望为 0.2 年的指数分布.

用 X_j 表示 C_j 时刻发生的损失在 T_j 时刻的赔付额. 假设 X_j 与 C_j 是独立的 (索赔量不依赖于它在一年中发生的时间), 但是 X_j 和 L_j 是正相关的 (在这个例子的续中将假定一个特殊的分布模型). 这是合理的, 因为更大的损失量需要用更多的时间来结案.

最后，用 V_t 表示在 t 年之后累积到 1 个货币单位现在投入的货币量. 它与所有的 X_j, C_j, L_j 相独立. 但显然, 对于 $s \neq t$, V_s 与 V_t 不独立.

最终有

$$S = \sum_{j=1}^{N} X_j V_{T_j},$$

其中 $N = \max_{C_j<1}\{j\}$. 在推导这个随机变量的过程中充分体现了各种其他变量之间的相关关系. □

例 17.7(含投保人最大自留额的情形) 假设对每一笔损失都考虑一个免赔额 d, 但是, 在一年中, 投保人自己的支付不会超过 u. 试建立保险人的总赔付模型.

解 用 X_j 表示第 j 笔损失. 这里 j 的次序没有关系. 用 $W_j = X_j \wedge d$ 表示有免赔时投保人自己的支出, 用 $Y_j = X_j - W_j$ 表示保险人的赔付额. $R = W_1 + \cdots + W_N$ 表示没有支付上限时投保人的总支付额, 则投保人的实际支出为 $R_u = R \wedge u$. 令 $S = X_1 + \cdots + X_N$ 表示实际的总损失, 则保险人的总赔付是 $T = S - R_u$. 注意 S 和 R_u 的分布都是基于独立同分布的索赔额分布, 可以用前面介绍的解析方法来得到它们的分布. 但是因为存在相依性, 不能通过合并分布的方法生成 T 的分布. 也不存在一种办法将 T 表示成独立同分布随机变量 Y_j 的随机和的形式. 在年初的时候, 看起来 T 可以表示成独立同分布的 Y_j 的和, 但是在某些点当投保人的总支付上界达到时 Y_j 将被 X_j 所代替. □

17.2.3 两个例子的模拟分析

现在用模拟的方法来完成上面的两个例子. 虽然模型的选择是任意的, 但是假定这里的模型是由前面介绍的技术方法通过仔细的估计过程得到的.

例 17.8(续例 17.6) 假设模型具有以下条件. 赔付额 (X_j) 服从参数为 $\alpha = 3, \theta = 1\,000$ 的 Pareto 分布. 从索赔发生到赔付结案的时间 (L_j) 服从参数为 $\tau = 1.5$ 和 $\theta = \ln(X_j)/6$ 的 Weibull 分布, 这里让尺度参数依赖于损失量建立了两个变量之间的相依关系. 关于贴现因子的模型假设如下, 当 $t > s$, $[\ln(V_s/V_t)]/(t-s)$ 服从均值为 0.06, 方差为 $0.000\,4(t-s)$ 的正态分布. 不需要对损失个数假定一个模型, 因为已经用损失发生间隔的模型作为替代. 试用随机模拟计算这时的总赔付的期望现值.

解 这里将详细介绍每一次模拟的细节, 具体说明计算过程. 首先, 生成一列独立的共同分布为指数分布的损失间隔时间, 直到这些值的和超过 1(这样就得到了一年内各次索赔的时间). 每个值由伪均匀分布的随机数按下式生成

$$u = 1 - \mathrm{e}^{-5x},$$

即

$$x = -0.2\ln(1-u).$$

第一次模拟的均匀分布伪随机数和相应的 x 值为：(0.253 73, 0.058 5), (0.467 50, 0.126 0), (0.237 09, 0.054 1), (0.757 80, 0.283 6) 以及 (0.966 42, 0.678 8). 此时的 x 值之和为 1.201 0, 因此一年中有 4 次损失按间隔 $c_1 = 0.058\ 5$、$c_2 = 0.184\ 5$、$c_3 = 0.238\ 6$ 和 $c_4 = 0.522\ 2$ 发生.

4 次损失的金额由 Pareto 分布函数的逆得到. 即,

$$x = 1\ 000[(1-u)^{-1/3} - 1].$$

4 个均匀分布伪随机数为 0.717 86、0.477 79、0.610 84 和 0.685 79. 由此得到的损失量为 $x_1 = 524.68$、$x_2 = 241.80$、$x_3 = 369.70$ 和 $x_4 = 470.93$.

已知索赔发生到赔付结案的时间服从 Weibull 分布. 求解的方程为

$$u = 1 - \mathrm{e}^{-[6l/\ln(x)]^{1.5}},$$

其中 x 代表损失量. 解得时间间隔 l 为

$$l = \frac{1}{6}\ln(x)[-\ln(1-u)]^{2/3}.$$

关于第一个间隔, 由 $u = 0.233\ 76$, 因此有

$$l_1 = \frac{1}{6}\ln(524.68)[-\ln 0.766\ 24]^{2/3} = 0.432\ 0.$$

类似地, 对于接下来 3 个 u 值 0.857 99, 0.129 51, 0.720 85, 有 $l_2 = 1.428\ 6$, $l_3 = 0.264\ 0$, $l_4 = 1.206\ 8$. 损失最终的支付时间是 c_j 与 l_j 之和, 分别为 $t_1 = 0.490\ 5$, $t_2 = 1.613\ 1$, $t_3 = 0.502\ 6$, $t_4 = 1.729\ 0$.

最后生成贴现因子. 它们必须按照 t_j 的上升顺序生成, 因此首先生成 $v_{0.4905}$. 从一个均值为 0.06, 方差为 0.000 4(0.490 5) = 0.000 196 2 的正态随机变量开始, 经过逆变换由模拟值 $0.059\ 2 = [\ln(1/v_{0.4905})]/0.490\ 5$ 得到 $v_{0.4905} = 0.971\ 4$. 注意第一个值为 $s = 0$ 以及 $v_0 = 1$. 第二个值是均值为 0.06 方差为 (0.502 6 − 0.490 5)(0.000 4) = 0.000 004 84 的正态变量, 模拟值为

$$0.060\ 4 = \frac{\ln(0.971\ 4/v_{0.502\ 6})}{0.012\ 1}, v_{0.502\ 6} = 0.970\ 7.$$

对于随后的两个索赔, 有

$$0.076\ 8 = \frac{\ln(0.970\ 7/v_{1.613\ 1})}{1.110\ 5}, v_{1.613\ 1} = 0.891\ 3.$$

$$0.062\ 8 = \frac{\ln(0.891\ 3/v_{1.729\ 0})}{0.115\ 9}, v_{1.729\ 0} = 0.884\ 8.$$

现在可以计算总现值的第一个模拟值, 为

$$\begin{aligned} s_1 &= 524.68(0.971\ 4) + 241.80(0.891\ 3) + 369.70(0.970\ 7) + 470.93(0.884\ 8) \\ &= 1\ 500.74. \end{aligned}$$

重复这个过程直至样本均值以 95%置信水平位于真实值的 1%左右. 总共需要 26 944 次模拟, 生成的样本均值为 2 299.16. □

例 17.9(续例 17.7) **这时的免赔额为 250, 赔付上界为 $u = 1\ 000$. 假设损失次数服从参数为 $r = 3, \beta = 2$ 的负二项分布. 进一步假设个体损失额服从参数为 $\tau = 2, \theta = 600$ 的 Weibull 分布. 试给出保险人损失的 95%分位数.**

解 为了生成负二项分布的索赔次数, 我们需要参照负二项分布的分布函数. 它没有解析的形式, 但是可以构造分布函数表, 正如表 17-2 所示. 要得到一年中的损失数, 首先生成一个均匀分布伪随机数, 例如 $u = 0.475\ 15$, 然后确定表中比 0.475 15 大的最小的分布函数值, 生成的随机数即为该值左边的对应值. 在这个例子中, 第一次随机模拟数为 $n = 5$ 的损失.

表 17-2 负二项分布函数值

n	$F_N(n)$	n	$F_N(n)$
0	0.037 04	8	0.765 89
1	0.111 11	9	0.818 88
2	0.209 88	10	0.861 27
3	0.319 62	11	0.894 67
4	0.429 36	12	0.920 64
5	0.531 78	13	0.940 62
6	0.622 82	14	0.955 85
7	0.700 86	15	0.967 35

5 次损失的索赔量分别由 Weibull 分布得到. 由分布函数取逆得到

$$x = 600[-\ln(1-u)]^{1/2}.$$

5 个随机数为：544.04, 453.67, 217.87, 681.98, 449.83. 总损失为 2 347.39. 因此, 投保人需要支付 $250.00 + 250.00 + 217.87 + 250.00 + 250.00 = 1\ 217.87$ 元, 但是投保人的支出上界为 1 000, 因此保险人赔付额的第一个模拟值为 1 347.39.

若目标为 95%置信水平的 95%分位点在真值的 2%左右, 需要 11 476 次模拟, 得到的 95%分位点估计值为 6 668.18. □

17.2.4 统计分析

在数据分析时, 随机模拟也可以提供很多帮助. 这里将讨论其中的两个方面, 这两种方法都与对统计过程的评价有关. 第一是确定假设检验的 p 值 (或临界值). 第

二是估计统计量的均方误差. 我们从假设检验开始.

例 17.10 假设损失服从对数正态分布. 现有 100 个观测并且 Kolmogorov-Smirnov 检验统计量为 0.062 72. 确定该检验的 p 值, 首先考虑零假设是参数为 $\mu = 7, \sigma = 1$ 的对数正态分布, 然后是参数未知的对数正态分布.

解 首先考虑参数值已知的零假设, 每次模拟包含由对数正态分布产生的 100 个观测值, 并计算 Kolmogorov-Smirnov 统计量. p 值由模拟中检验统计量超过 0.062 72 的比例进行估计. 1 000 次模拟的 p 值估计为 0.836.

若参数未知, 并未明确应采用的对数正态分布. 由实际观测值估计得到: $\hat{\mu} = 7.220\ 1$ 和 $\hat{\sigma} = 0.808\ 93$. 这将作为每次模拟的基础, 唯一的改变是当得到模拟观测值之后, 将与模拟数据的最大似然估计的对数正态分布进行比较. 对于 1 000 次模拟, 检验统计量超过 0.062 72 的共计 491 次, 因此估计的 p 值为 0.491.

正如 13.4.1 节所述, 参数已知和未知对统计检验的解释有显著的不同. □

当检验假设时, p 值和显著性水平是在零假设成立的条件下计算的, 在其他情况下没有可采用的已知分布. 对于这些情况, 可以借助被称为自助法的技术 (要得到这个方法的详细介绍, 请见 [33]), 这种方法的本质是将由数据得到的经验分布作为总体来生成模拟值. 理论结果显示至少自助法得到的估计会渐进地收敛到真实值. 这样做是合理的, 因为当样本容量增大时, 经验分布将会越来越接近真实分布. 下面的例子将具体说明自助法的使用, 并且表明至少在所给的例子中它给出了一个合理的答案.

例 17.11 来自总体的一个容量为 3 的样本 (可以含重复观测) 的值是 2,3,7. 确定以样本均值作为总体均值估计的均方误差的自助法估计.

解 首先自助法假设从总体中抽取 2,3,7 的概率均为 1/3, 因此分布的均值为 4. 基于这个总体共有 27 个容量为 3 的样本可以抽取. 样本均值分别为 2(2,2,2 抽到的概率是 1/27), 7/3(2,2,3, 2,3,3 或 3,2,2, 抽到的概率均为 3/27) 等等, 最大均值为 7, 抽到的概率是 1/27. 因此, 均方误差为

$$(2-4)^2(1/27) + \left(\frac{7}{3} - 4\right)^2 (3/27) + \cdots + (7-4)^2(1/27) = \frac{14}{9}.$$

另外, 通常的方法是基于样本均值是无偏估计, 则有

$$\mathrm{MSE}(\bar{X}) = \mathrm{Var}(\bar{X}) = \sigma^2/n.$$

当方差未知时, 一个合理的选择是用样本方差代替. 分母为 n, 在本例中估计的均方误差为

$$\frac{\frac{1}{3}[(2-4)^2 + (3-4)^2 + (7-4)^2]}{3} = \frac{14}{9},$$

与自助法得到的估计相同. □

在许多情况下, 均方误差的计算不会这样容易, 因此自助法变得非常有用. 虽然对于本例并不需要进行模拟, 因为容量为 3 的样本最多生成 27 个可能的自助值. 但当样本容量超过 6 时, 要想列举出所有的情况是不容易的. 这时, 由经验分布模拟观测值变成了唯一可行的选择.

例 17.12 由例 11.3 得到生存时间的经验模型, 已知死亡发生在时刻 0.8, 2.9, 3.1, 4.0, 4.1, 4.8 的概率分别为 0.033 3, 0.074 4, 0.034 3, 0.066 0, 0.034 4, 0.036 1, 最后剩余的概率 0.721 5 是某人从现在起生存 5 年以上的概率. 死亡当时赔付 1 000 元的 5 年定期寿险的期望现值为

$$1\,000(0.033\,3v^{0.8}+\cdots+0.036\,1v^{4.8})=223.01,$$

其中 $v=1.07^{-1}$. 用自助法模拟 10 000 个样本估计这个估计量的均方误差.

解 Efron[31] 给出了一种使用 Kaplan-Meier 估计进行自助法模拟的方法. 模拟值不是来自经验分布 (由 Kaplan-Meier 估计给出), 而是来自样本. 本例对每个原始观测赋于 1/40 的概率, 那么每个自助观测都是一个伴随删失或未删失值的左截断点. 当记录了 40 个这样的观测后, 对这些自助法样本构造 Kaplan-Meier 估计然后计算所关心的值. 这样做相对简单, 因为这种自助估计只对 6 个原始点赋概率值, 10 000 次模拟很快就完成了, 得到的均值为 222.05, 均方误差为 4,119. Efron 还注意到自助法估计 $\hat{S}(t)$ 的方差与 Greenwood 估计渐近相等, 这使得两个方法都变得更加可信. □

习题

17.7* 现有对某城市的清除积雪费用保险, 期限为 4 个冬季的月份, 每个月为 10 000 的免赔额, 每个月的费用是独立的并且服从 $\mu=15\ 000$ 和 $\sigma=2\ 000$ 的正态分布. 由逆变换法模拟得到每个月的费用. 现有一年总费用的某次模拟, 4 个均匀分布的伪随机数为: 0.539 8, 0.115 1, 0.001 3, 0.788 1. 计算这次模拟的保险人年总支付.

17.8* 已知期末时某股票的价格是期初的 X 倍, 其中 X 服从 $\mu=0.01$ 和 $\sigma=0.02$ 的对数正态分布. 0 时刻的价格是 100. 试用逆变换方法模拟价格变化. 若第 1 期和第 2 期的伪均匀分布随机数分别是 0.158 7 和 0.933 2. 确定前两期期末价格的模拟值.

17.9* 现有 100 名 70 岁老人的一年定期寿险. 每人下一年的身故概率是 0.033 18 并且是独立的. 因此身故数服从参数为 $m=100$, $q=0.033\ 18$ 的二项分布. 用逆变换法确定下一年的死亡人数, 设均匀分布的伪随机数是 $u=0.18$.

17.10* 已知某盈余过程按照年索赔率为 2 的 Poisson 过程发生索赔, 因此索赔发生间隔服从 $\theta=2$ 的指数分布. 已知索赔量服从参数为 $\alpha=2$ 和 $\theta=1\ 000$ 的 Pareto 分布. 初始盈余为 2 000 并且保费按照 2 200 的速率收取. 破产在盈余为负的时刻发生, 之后没有进一步的保费收入和索赔支出. 所有模拟用逆变换法完成. 用 0.83, 0.54, 0.48, 0.14 作为伪均匀分布随机数产生索赔间隔时间. 用 0.89, 0.36, 0.70, 0.61 产生索赔量. 试确定时刻 1 的盈余.

17.11* 现有样本容量为 2 的随机样本 1 和 3. 考虑用 $[(X_1-\bar{X})^2+(X_2-\bar{X})^2]/2$ 估计总体方差. 试用自助法估计这个估计量的均方误差.

17.12 现有 (0,10) 上均匀分布的容量为 3 的样本：2, 4, 7.

(a) 对于 "数据为 (0,10) 上均匀分布" 的零假设, 计算 Kolmogorov-Smirnov 检验统计量.

(b) 对 (0,10) 均匀分布模拟 10 000 个容量为 3 的样本, 并对每次模拟计算 Kolmogorov-Smirnov 检验统计量, 其中等于或超过 (a) 中结果的比例作为 p 值的估计.

17.13 现有 $(0,\theta)$ 上均匀分布的容量为 3 的样本：2, 4, 7. 考虑如下关于 θ 的估计量

$$\hat{\theta}=\frac{4}{3}\max(x_1,x_2,x_3).$$

由例 9.15 知该无偏估计量的均方误差为 $\theta^2/15$.

(a) 以 θ 的估计代替 θ 来估计均方误差.

(b) 用自助法估计统计量的方差. (用自助法是不能够估计均方误差的, 因为由经验分布无法得到 θ 的真实值, 但是可以得到估计量的期望值.)

附录A　连续分布函数

A.1　引　　言

下面对一些常见的分布进行介绍. 首先是数学预备知识, 说明各种量的计算.

不完全 gamma 函数①的定义为

$$\Gamma(\alpha;x)=\frac{1}{\Gamma(\alpha)}\int_0^x t^{\alpha-1}\mathrm{e}^{-t}\mathrm{d}t,\quad \alpha>0,x>0,$$

其中

$$\Gamma(\alpha)=\int_0^\infty t^{\alpha-1}\mathrm{e}^{-t}\mathrm{d}t,\quad \alpha>0.$$

又定义

$$G(\alpha;x)=\int_x^\infty t^{\alpha-1}\mathrm{e}^{-t}\mathrm{d}t,\quad x>0.$$

有时需要对非正数 α 计算这个积分, 利用分部积分可得到关系式

$$G(\alpha;x)=-\frac{x^\alpha \mathrm{e}^{-x}}{\alpha}+\frac{1}{\alpha}G(\alpha+1;x).$$

反复进行分部积分, 直到 G 的第一个自变量为正数 $\alpha+k$, 然后可通过下式求值

$$G(\alpha+k;x)=\Gamma(\alpha+k)[1-\Gamma(\alpha+k;x)].$$

然而, 如果 α 是一个负整数或零, 需要定义 $G(0;x)$ 的值

$$G(0;x)=\int_x^\infty t^{-1}\mathrm{e}^{-t}\mathrm{d}t=E_1(x),$$

称之为**指数积分**. 这个积分的级数展开为

$$E_1(x)=-0.577\,215\,664\,901\,53-\ln x-\sum_{n=1}^\infty\frac{(-1)^n x^n}{n(n!)}.$$

当 α 为正整数时, 不完全 gamma 函数可直接通过下面的定理进行求值.

---

① 很多参考书目, 比如 [3], 将这个积分表示为 $P(\alpha,x)$, 并定义 $\Gamma(\alpha,x)=\int_x^\infty t^{\alpha-1}\mathrm{e}^{-t}\mathrm{d}t$. 注意这个定义没有除以 $\Gamma(\alpha)$ 进行标准化, 在使用计算机程序计算不完全 gamma 函数时, 要首先确定那里的具体定义.

定理 A.1 对整数 α, 有

$$\Gamma(\alpha;x)=1-\sum_{j=0}^{\alpha-1}\frac{x^j\mathrm{e}^{-x}}{j!}.$$

证明 对 $\alpha=1$, $\Gamma(1;x)=\int_0^x \mathrm{e}^{-t}\mathrm{d}t=1-\mathrm{e}^{-x}$, 所以定理在这种情况下成立. 定理的证明可通过归纳法完成, 假设定理对 $\alpha=1,\cdots,n$ 成立, 则

$$\begin{aligned}\Gamma(n+1;x) &= \frac{1}{n!}\int_0^x t^n\mathrm{e}^{-t}\mathrm{d}t\\ &= \frac{1}{n!}\left(-t^n\mathrm{e}^{-t}\Big|_0^x+\int_0^x nt^{n-1}\mathrm{e}^{-t}\mathrm{d}t\right)\\ &= \frac{1}{n!}(-x^n\mathrm{e}^{-x})+\Gamma(n;x)\\ &= -\frac{x^n\mathrm{e}^{-x}}{n!}+1-\sum_{j=0}^{n-1}\frac{x^j\mathrm{e}^{-x}}{j!}\\ &= 1-\sum_{j=0}^{n}\frac{x^j\mathrm{e}^{-x}}{j!}.\end{aligned}$$

不完全 beta 函数的定义为

$$\beta(a,b;x)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^x t^{a-1}(1-t)^{b-1}\mathrm{d}t,\quad a>0,\quad b>0,\quad 0<x<1,$$

而且, 当 $b<0$(但是 $a>1+\lfloor -b\rfloor$) 时, 反复利用分部积分可得到

$$\begin{aligned}\Gamma(a)\Gamma(b)\beta(a,b;x) = &-\Gamma(a+b)\left[\frac{x^{a-1}(1-x)^b}{b}\right.\\ &+\frac{(a-1)x^{a-2}(1-x)^{b+1}}{b(b+1)}+\cdots\\ &\left.+\frac{(a-1)\cdots(a-r)x^{a-r-1}(1-x)^{b+r}}{b(b+1)\cdots(b+r)}\right]\\ &+\frac{(a-1)\cdots(a-r-1)}{b(b+1)\cdots(b+r)}\Gamma(a-r-1)\cdot\\ &\times\Gamma(b+r+1)\beta(a-r-1,b+r+1;x),\end{aligned}$$

其中 r 是满足 $b+r+1>0$ 的最小整数, 第一个自变量必须是正的, 即 $a-r-1>0$.

关于不完全 gamma 函数和不完全 beta 函数的数值近似在很多统计计算和电子表程序中都有所实现, 因为它们也是 gamma 分布和 beta 分布的分布函数. 下面的近似取自参考文献 [3]. 计算不完全 gamma 函数时, 对较小的 x 和较大的 x 将分别使用不同的方程, 见参考文献 [107]. 该参考书还给出了计算这些表达式的计算机子程序. 特别地, 它提供了一种计算连分式的有效方法.

对 $x \leqslant \alpha+1$, 使用级数展开式

$$\Gamma(\alpha;x)=\frac{x^{\alpha}\mathrm{e}^{-x}}{\Gamma(\alpha)}\sum_{n=0}^{\infty}\frac{x^n}{\alpha(\alpha+1)\cdots(\alpha+n)}.$$

当 $x>\alpha+1$ 时, 使用连分式展开式

$$1-\Gamma(\alpha;x)=\frac{x^{\alpha}\mathrm{e}^{-x}}{\Gamma(\alpha)}\cfrac{1}{x+\cfrac{1-\alpha}{1+\cfrac{1}{x+\cfrac{2-\alpha}{1+\cfrac{2}{x+\cdots}}}}}.$$

不完全 gamma 函数也可以用来计算标准正态分布的累积概率. 令 $\Phi(z)=\Pr(Z\leqslant z)$,$Z$ 服从标准正态分布. 当 $z\geqslant 0$ 时, $\Phi(z)=0.5+\Gamma(0.5;z^2/2)/2$, 当 $z<0$ 时, $\Phi(z)=1-\Phi(-z)$.

不完全 beta 函数可通过级数展开式计算

$$\beta(a,b;x)=\frac{\Gamma(a+b)x^a(1-x)^b}{a\Gamma(a)\Gamma(b)}\times\left[1+\sum_{n=0}^{\infty}\frac{(a+b)(a+b+1)\cdots(a+b+n)}{(a+1)(a+2)\cdots(a+n+1)}x^{n+1}\right].$$

还可以用如下近似计算 gamma 函数

$$\begin{aligned}\ln\Gamma(\alpha)\doteq{}&\left(\alpha-\frac{1}{2}\right)\ln\alpha-\alpha+\frac{\ln(2\pi)}{2}\\&+\frac{1}{12\alpha}-\frac{1}{360\alpha^3}+\frac{1}{1\,260\alpha^5}-\frac{1}{1\,680\alpha^7}+\frac{1}{1\,188\alpha^9}-\frac{691}{360\,360\alpha^{11}}\\&+\frac{1}{156\alpha^{13}}-\frac{3\,617}{122\,400\alpha^{15}}+\frac{43\,867}{244\,188\alpha^{17}}-\frac{174\,611}{125\,400\alpha^{19}}.\end{aligned}$$

当 α 大于 10 时, 这个近似的误差小于 10^{-19}, α 小于 10 时, 利用如下关系式

$$\ln\Gamma(\alpha)=\ln\Gamma(\alpha+1)-\ln\alpha.$$

以下对概率函数的介绍均按照如下方式进行. 首先给出含参数的分布名称, 如果有其他的名称将在括号中标注. 然后是密度函数 $f(x)$ 和分布函数 $F(x)$, 还对很多分布给出了方程的初值. 在同一分布类中, 概率函数按照参数个数递减的顺序介绍. 使用的希腊字母是统一的, 如果概率函数中未使用任何希腊字母, 则意味着这

个分布为多个参数中的一个特例, 而缺失的参数等于 1. 除非特别说明, 所有参数都取正值.

除了两个分布以外, 都可以通过缩放尺度系数 θ 实现对变量的缩放. 也就是说, 如果已知 X 具有某个分布, 则 cX 具有与之相同的分布类型, 只需将 θ 改为 $c\theta$ 其他参数均保持不变. 对于对数正态分布, μ 改为 $\mu+\ln(c)$ 而 σ 不变, 而对逆 Gaussian 分布, μ 和 θ 都要乘以 c.

我们还对很多概率函数都给出了建议的初值, 不一定是最好的估计, 只是一个开始反复迭代的初值, 可以用来最大化似然函数或其他目标函数. 这些值可以通过矩方法或分位点匹配的方法得到, 相关的量为

$$\text{矩方法：} m = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad t = \frac{1}{n}\sum_{i=1}^{n} x_i^2,$$
$$\text{分位点匹配：} p = 25\%\ \text{分位点}, q = 75\%\ \text{分位点}.$$

对于分组数据、截断数据或删失数据, 只能近似得到上述这些量. 因为目的是得到初始值, 而不是有效的估计量, 通常不必进行修正. 对于含有 3 个或 4 个参数的概率分布, 初值可通过对特殊情况的估计而得到, 然后令新的参数等于 1. 一种通用的初值原则 (当其他方法都失效时) 是令尺度参数 (θ) 等于均值, 而其他参数等于 2.

这里列举的所有概率分布 (还有其他很多) 在 [73] 中都有非常详细的讨论. 很多情况下, 也给出了最大似然估计量.

A.2　转换 beta 分布族

A.2.1　四参数分布

转换 beta 分布 —α, θ, γ, τ(第二类广义 beta 分布, Pearson Type Ⅵ)

$$\begin{aligned}
f(x) &= \frac{\Gamma(\alpha+\tau)}{\Gamma(\alpha)\Gamma(\tau)}\frac{\gamma(x/\theta)^{\gamma\tau}}{x[1+(x/\theta)^{\gamma}]^{\alpha+\tau}},\\
F(x) &= \beta(\tau,\alpha;u), \quad u = \frac{(x/\theta)^{\gamma}}{1+(x/\theta)^{\gamma}},\\
\mathrm{E}[X^k] &= \frac{\theta^k\Gamma(\tau+k/\gamma)\Gamma(\alpha-k/\gamma)}{\Gamma(\alpha)\Gamma(\tau)}, \quad -\tau\gamma < k < \alpha\gamma,\\
\mathrm{E}[(X\wedge x)^k] &= \frac{\theta^k\Gamma(\tau+k/\gamma)\Gamma(\alpha-k/\gamma)}{\Gamma(\alpha)\Gamma(\tau)}\beta(\tau+k/\gamma,\alpha-k/\gamma;u)\\
&\quad + x^k[1-F(x)], \qquad k > -\tau\gamma,
\end{aligned}$$

$$众数 = \theta\left(\frac{\tau\gamma-1}{\alpha\gamma+1}\right)^{1/\gamma}, \quad \tau\gamma>1, \text{ 其他 } 0.$$

A.2.2 三参数分布

广义 Pareto 分布 —α, θ, τ. (第二类 beta 分布)

$$\begin{aligned}
f(x) &= \frac{\Gamma(\alpha+\tau)}{\Gamma(\alpha)\Gamma(\tau)}\frac{\theta^\alpha x^{\tau-1}}{(x+\theta)^{\alpha+\tau}}, \\
F(x) &= \beta(\tau,\alpha;u), \quad u=\frac{x}{x+\theta}, \\
\mathrm{E}[X^k] &= \frac{\theta^k\Gamma(\tau+k)\Gamma(\alpha-k)}{\Gamma(\alpha)\Gamma(\tau)}, \quad -\tau<k<\alpha, \\
\mathrm{E}[X^k] &= \frac{\theta^k\tau(\tau+1)\cdots(\tau+k-1)}{(\alpha-1)\cdots(\alpha-k)} \text{ 如果 } k \text{ 为整数}, \\
\mathrm{E}[(X\wedge x)^k] &= \frac{\theta^k\Gamma(\tau+k)\Gamma(\alpha-k)}{\Gamma(\alpha)\Gamma(\tau)}\beta(\tau+k,\alpha-k;u), \\
&\quad + x^k[1-F(x)], \quad k>-\tau, \\
众数 &= \theta\frac{\tau-1}{\alpha+1}, \quad \tau>1, \text{ 其他 } 0.
\end{aligned}$$

Burr 分布 —α, θ, γ. (Burr Type Ⅻ, Singh-Maddala)

$$\begin{aligned}
f(x) &= \frac{\alpha\gamma(x/\theta)^\gamma}{x[1+(x/\theta)^\gamma]^{\alpha+1}}, \\
F(x) &= 1-u^\alpha, \quad u=\frac{1}{1+(x/\theta)^\gamma}, \\
\mathrm{E}[X^k] &= \frac{\theta^k\Gamma(1+k/\gamma)\Gamma(\alpha-k/\gamma)}{\Gamma(\alpha)}, \quad -\gamma<k<\alpha\gamma, \\
\mathrm{E}[(X\wedge x)^k] &= \frac{\theta^k\Gamma(1+k/\gamma)\Gamma(\alpha-k/\gamma)}{\Gamma(\alpha)}\beta(1+k/\gamma,\alpha-k/\gamma;1-u) \\
&\quad + x^k u^\alpha, \qquad k>-\gamma, \\
众数 &= \theta\left(\frac{\gamma-1}{\alpha\gamma+1}\right)^{1/\gamma}, \quad \gamma>1, \text{ 其他 } 0.
\end{aligned}$$

逆 Burr 分布 —τ, θ, γ. (Dagum)

$$\begin{aligned}
f(x) &= \frac{\tau\gamma(x/\theta)^{\gamma\tau}}{x[1+(x/\theta)^\gamma]^{\tau+1}}, \\
F(x) &= u^\tau, \quad u=\frac{(x/\theta)^\gamma}{1+(x/\theta)^\gamma}, \\
\mathrm{E}[X^k] &= \frac{\theta^k\Gamma(\tau+k/\gamma)\Gamma(1-k/\gamma)}{\Gamma(\tau)}, \quad -\tau\gamma<k<\gamma,
\end{aligned}$$

$$\mathrm{E}[(X\wedge x)^k]=\frac{\theta^k\Gamma(\tau+k/\gamma)\Gamma(1-k/\gamma)}{\Gamma(\tau)}\beta(\tau+k/\gamma,1-k/\gamma;u)$$
$$+x^k[1-u^\tau],\quad k>-\tau\gamma,$$
$$\text{众数}=\theta\left(\frac{\tau\gamma-1}{\gamma+1}\right)^{1/\gamma},\quad \tau\gamma>1,\ \text{其他}\,0.$$

A.2.3 两参数分布

Pareto 分布 —α, θ(Pareto Type Ⅱ, Lomax)

$$f(x)=\frac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}},$$
$$F(x)=1-\left(\frac{\theta}{x+\theta}\right)^\alpha,$$
$$\mathrm{E}[X^k]=\frac{\theta^k\Gamma(k+1)\Gamma(\alpha-k)}{\Gamma(\alpha)},\quad -1<k<\alpha,$$
$$\mathrm{E}[X^k]=\frac{\theta^k k!}{(\alpha-1)\cdots(\alpha-k)}\quad \text{如果}\,k\,\text{为整数},$$
$$\mathrm{E}[X\wedge x]=\frac{\theta}{\alpha-1}\left[1-\left(\frac{\theta}{x+\theta}\right)^{\alpha-1}\right],\quad \alpha\neq 1,$$
$$\mathrm{E}[X\wedge x]=-\theta\ln\left(\frac{\theta}{x+\theta}\right),\quad \alpha=1,$$
$$\mathrm{E}[(X\wedge x)^k]=\frac{\theta^k\Gamma(k+1)\Gamma(\alpha-k)}{\Gamma(\alpha)}\beta[k+1,\alpha-k;x/(x+\theta)]$$
$$+x^k\left(\frac{\theta}{x+\theta}\right)^\alpha,\quad \text{所有}\,k,$$
$$\text{众数}=0,$$
$$\hat{\alpha}=2\frac{t-m^2}{t-2m^2},\quad \hat{\theta}=\frac{mt}{t-2m^2}.$$

逆 Pareto 分布 —τ, θ.

$$f(x)=\frac{\tau\theta x^{\tau-1}}{(x+\theta)^{\tau+1}},$$
$$F(x)=\left(\frac{x}{x+\theta}\right)^\tau,$$
$$\mathrm{E}[X^k]=\frac{\theta^k\Gamma(\tau+k)\Gamma(1-k)}{\Gamma(\tau)},\quad -\tau<k<1,$$
$$\mathrm{E}[X^k]=\frac{\theta^k(-k)!}{(\tau-1)\cdots(\tau+k)}\quad \text{当}\,k\,\text{为负整数时},$$

$$\mathrm{E}[(X\wedge x)^k]=\theta^k\tau\int_0^{x/(x+\theta)}y^{\tau+k-1}(1-y)^{-k}dy+x^k\left[1-\left(\frac{x}{x+\theta}\right)^{\tau}\right],\quad k>-\tau,$$

$$\text{众数}=\theta\frac{\tau-1}{2},\quad \tau>1,\ \text{其他 } 0.$$

Loglogistic 分布 —γ, θ(Fisk)

$$\begin{aligned}
f(x)&=\frac{\gamma(x/\theta)^{\gamma}}{x[1+(x/\theta)^{\gamma}]^2},\\
F(x)&=u,\quad u=\frac{(x/\theta)^{\gamma}}{1+(x/\theta)^{\gamma}},\\
\mathrm{E}[X^k]&=\theta^k\Gamma(1+k/\gamma)\Gamma(1-k/\gamma),\quad -\gamma<k<\gamma,\\
\mathrm{E}[(X\wedge x)^k]&=\theta^x\Gamma(1+k/\gamma)\Gamma(1-k/\gamma)\beta(1+k/\gamma,1-k/\gamma;u)\\
&\quad+x^k(1-u),\quad k>-\gamma,\\
\text{众数}&=\theta\left(\frac{\gamma-1}{\gamma+1}\right)^{1/\gamma},\quad \gamma>1,\ \text{其他 } 0,\\
\hat{\gamma}&=\frac{2\ln(3)}{\ln(q)-\ln(p)},\quad \hat{\theta}=\exp\left(\frac{\ln(q)+\ln(p)}{2}\right).
\end{aligned}$$

Paralogistic—α, θ. 这是 $\gamma=\alpha$ 时的 Burr 分布.

$$\begin{aligned}
f(x)&=\frac{\alpha^2(x/\theta)^{\alpha}}{x[1+(x/\theta)^{\alpha}]^{\alpha+1}},\\
F(x)&=1-u^{\alpha},\quad u=\frac{1}{1+(x/\theta)^{\alpha}},\\
\mathrm{E}[X^k]&=\frac{\theta^k\Gamma(1+k/\alpha)\Gamma(\alpha-k/\alpha)}{\Gamma(\alpha)},\quad -\alpha<k<\alpha^2,\\
\mathrm{E}[(X\wedge x)^k]&=\frac{\theta^k\Gamma(1+k/\alpha)\Gamma(\alpha-k/\alpha)}{\Gamma(\alpha)}\beta(1+k/\alpha,\alpha-k/\alpha;1-u)\\
&\quad+x^ku^{\alpha},\quad k>-\alpha,\\
\text{众数}&=\theta\left(\frac{\alpha-1}{\alpha^2+1}\right)^{1/\alpha},\quad \alpha>1,\ \text{其他 } 0
\end{aligned}$$

可以使用 loglogistic 分布 (用 γ 代替 α) 或 Pareto 分布 (使用 α) 的估计值为初值.

逆 paralogistic—τ, θ. 这是 $\gamma=\tau$ 时的逆 Burr 分布.

$$\begin{aligned}
f(x)&=\frac{\tau^2(x/\theta)^{\tau^2}}{x[1+(x/\theta)^{\tau}]^{\tau+1}},\\
F(x)&=u^{\tau},\quad u=\frac{(x/\theta)^{\tau}}{1+(x/\theta)^{\tau}},
\end{aligned}$$

$$\mathrm{E}[X^k] = \frac{\theta^k\Gamma(\tau+k/\tau)\Gamma(1-k/\tau)}{\Gamma(\tau)}, \quad -\tau^2 < k < \tau,$$

$$\mathrm{E}[(X\wedge x)^k] = \frac{\theta^k\Gamma(\tau+k/\tau)\Gamma(1-k/\tau)}{\Gamma(\tau)}\beta(\tau+k/\tau, 1-k/\tau; u) + x^k[1-u^\tau], \quad k > -\tau^2,$$

$$\text{众数} = \theta(\tau-1)^{1/\tau}, \quad \tau > 1, \text{ 其他 } 0.$$

可以使用 loglogistic 分布 (用 γ 代替 τ) 或逆 Pareto 分布 (使用 τ) 的估计值为初值.

A.3　转换 gamma 分布族

A.3.1　三参数分布

转换的 gamma 分布 —α, θ, τ(广义 gamma 分布)

$$f(x) = \frac{\tau u^\alpha \mathrm{e}^{-u}}{x\Gamma(\alpha)}, \quad u = (x/\theta)^\tau,$$

$$F(x) = \Gamma(\alpha; u),$$

$$\mathrm{E}[X^k] = \frac{\theta^k\Gamma(\alpha+k/\tau)}{\Gamma(\alpha)}, \quad k > -\alpha\tau,$$

$$\mathrm{E}[(X\wedge x)^k] = \frac{\theta^k\Gamma(\alpha+k/\tau)}{\Gamma(\alpha)}\Gamma(\alpha+k/\tau; u) + x^k[1-\Gamma(\alpha; u)], \quad k > -\alpha\tau,$$

$$\text{众数} = \theta\left(\frac{\alpha\tau-1}{\tau}\right)^{1/\tau}, \quad \alpha\tau > 1, \text{ 其他 } 0.$$

逆转换 gamma 分布 —α, θ, τ(广义逆 gamma 分布)

$$f(x) = \frac{\tau u^\alpha \mathrm{e}^{-u}}{x\Gamma(\alpha)}, \quad u = (\theta/x)^\tau,$$

$$F(x) = 1-\Gamma(\alpha; u),$$

$$\mathrm{E}[X^k] = \frac{\theta^k\Gamma(\alpha-k/\tau)}{\Gamma(\alpha)}, \quad k < \alpha\tau,$$

$$\mathrm{E}[(X\wedge x)^k] = \frac{\theta^k\Gamma(\alpha-k/\tau)}{\Gamma(\alpha)}[1-\Gamma(\alpha-k/\tau; u)] + x^k\Gamma(\alpha; u) = \frac{\theta^k G(\alpha-k/\tau; u)}{\Gamma(\alpha)} + x^k\Gamma(\alpha; u), \quad \text{所有 } k,$$

$$\text{众数} = \theta\left(\frac{\tau}{\alpha\tau+1}\right)^{1/\tau}.$$

A.3.2 两参数分布

Gamma 分布 —α, θ

$$
\begin{aligned}
f(x) &= \frac{(x/\theta)^{\alpha}\mathrm{e}^{-x/\theta}}{x\Gamma(\alpha)},\\
F(x) &= \Gamma(\alpha;x/\theta),\\
\mathrm{E}[X^k] &= \frac{\theta^k\Gamma(\alpha+k)}{\Gamma(\alpha)}, \quad k>-\alpha,\\
\mathrm{E}[X^k] &= \theta^k(\alpha+k-1)\cdots\alpha \quad \text{当 } k \text{ 为整数}\\
\mathrm{E}[(X\wedge x)^k] &= \frac{\theta^k\Gamma(\alpha+k)}{\Gamma(\alpha)}\Gamma(\alpha+k;x/\theta)+x^k[1-\Gamma(\alpha;x/\theta)], \quad k>-\alpha\\
\mathrm{E}[(X\wedge x)^k] &= \alpha(\alpha+1)\cdots(\alpha+k-1)\theta^k\Gamma(\alpha+k;x/\theta)\\
&\quad +x^k[1-\Gamma(\alpha;x/\theta)] \quad \text{当 } k \text{ 为整数},\\
M(t) &= (1-\theta t)^{-\alpha}, \quad t<1/\theta,\\
\text{众数} &= \theta(\alpha-1), \quad \alpha>1, \text{其他}\, 0,\\
\hat{\alpha} &= \frac{m^2}{t-m^2}, \quad \hat{\theta}=\frac{t-m^2}{m}.
\end{aligned}
$$

逆 gamma 分布 —α, θ(Vinci)

$$
\begin{aligned}
f(x) &= \frac{(\theta/x)^{\alpha}\mathrm{e}^{-\theta/x}}{x\Gamma(\alpha)},\\
F(x) &= 1-\Gamma(\alpha;\theta/x)\\
\mathrm{E}[X^k] &= \frac{\theta^k\Gamma(\alpha-k)}{\Gamma(\alpha)}, \quad k<\alpha,\\
\mathrm{E}[X^k] &= \frac{\theta^k}{(\alpha-1)\cdots(\alpha-k)} \quad \text{如果 } k \text{ 为整数}\\
\mathrm{E}[(X\wedge x)^k] &= \frac{\theta^k\Gamma(\alpha-k)}{\Gamma(\alpha)}[1-\Gamma(\alpha-k;\theta/x)]+x^k\Gamma(\alpha;\theta/x)\\
&= \frac{\theta^k G(\alpha-k;\theta/x)}{\Gamma(\alpha)}+x^k\Gamma(\alpha;\theta/x), \text{所有 } k,\\
\text{众数} &= \theta/(\alpha+1),\\
\hat{\alpha} &= \frac{2t-m^2}{t-m^2}, \quad \hat{\theta}=\frac{mt}{t-m^2}.
\end{aligned}
$$

Weibull 分布 —θ, τ

$$
f(x) = \frac{\tau(x/\theta)^{\tau}\mathrm{e}^{-(x/\theta)^{\tau}}}{x},
$$

$$
\begin{aligned}
F(x) &= 1 - \mathrm{e}^{-(x/\theta)^\tau}, \\
\mathrm{E}[X^k] &= \theta^k \Gamma(1 + k/\tau), \quad k > -\tau, \\
\mathrm{E}[(X \wedge x)^k] &= \theta^k \Gamma(1 + k/\tau)\Gamma[1 + k/\tau; (x/\theta)^\tau] + x^k \mathrm{e}^{-(x/\theta)^\tau}, \quad k > -\tau, \\
\text{众数} &= \theta \left(\frac{\tau - 1}{\tau}\right)^{1/\tau}, \quad \tau > 1,\ \text{其他}\ 0, \\
\hat{\theta} &= \exp\left(\frac{g\ln(p) - \ln(q)}{g - 1}\right), \quad g = \frac{\ln(\ln(4))}{\ln(\ln(4/3))}, \\
\hat{\tau} &= \frac{\ln(\ln(4))}{\ln(q) - \ln(\hat{\theta})}.
\end{aligned}
$$

逆 Weibull 分布 —θ, τ(log-Gompertz)

$$
\begin{aligned}
f(x) &= \frac{\tau(\theta/x)^\tau \mathrm{e}^{-(\theta/x)^\tau}}{x}, \\
F(x) &= \mathrm{e}^{-(\theta/x)^\tau}, \\
\mathrm{E}[X^k] &= \theta^k \Gamma(1 - k/\tau), \quad k < \tau, \\
\mathrm{E}[(X \wedge x)^k] &= \theta^k \Gamma(1 - k/\tau)\{1 - \Gamma[1 - k/\tau; (\theta/x)^\tau]\} + x^k \left[1 - \mathrm{e}^{-(\theta/x)^\tau}\right], \\
&= \theta^k G[1 - k/\tau; (\theta/x)^\tau] + x^k \left[1 - \mathrm{e}^{-(\theta/x)^\tau}\right], \quad \text{所有}\ k, \\
\text{众数} &= \theta \left(\frac{\tau}{\tau + 1}\right)^{1/\tau}, \\
\hat{\theta} &= \exp\left(\frac{g\ln(q) - \ln(p)}{g - 1}\right), \quad g = \frac{\ln(\ln(4))}{\ln(\ln(4/3))}, \\
\hat{\tau} &= \frac{\ln(\ln(4))}{\ln(\hat{\theta}) - \ln(p)}.
\end{aligned}
$$

A.3.3 单参数分布

指数分布 —θ

$$
\begin{aligned}
f(x) &= \frac{\mathrm{e}^{-x/\theta}}{\theta}, \\
F(x) &= 1 - \mathrm{e}^{-x/\theta}, \\
\mathrm{E}[X^k] &= \theta^k \Gamma(k + 1), \quad k > -1, \\
\mathrm{E}[X^k] &= \theta^k k! \quad \text{如果}\ k\ \text{为整数}, \\
\mathrm{E}[X \wedge x] &= \theta(1 - \mathrm{e}^{-x/\theta}), \\
\mathrm{E}[(X \wedge x)^k] &= \theta^k \Gamma(k + 1)\Gamma(k + 1; x/\theta) + x^k \mathrm{e}^{-x/\theta}, \quad k > -1, \\
\mathrm{E}[(X \wedge x)^k] &= \theta^k k! \Gamma(k + 1; x/\theta) + x^k \mathrm{e}^{-x/\theta} \quad \text{如果}\ k > -1\ \text{是一个整数},
\end{aligned}
$$

$$M(t)=(1-\theta t)^{-1},\quad t<1/\theta,$$
$$\text{众数}=0,$$
$$\hat{\theta}=m.$$

逆指数分布 —θ

$$f(x)=\frac{\theta \mathrm{e}^{-\theta/x}}{x^2},$$
$$F(x)=\mathrm{e}^{-\theta/x},$$
$$\mathrm{E}[X^k]=\theta^k\Gamma(1-k),\quad k<1,$$
$$\mathrm{E}[(X\wedge x)^k]=\theta^k G(1-k;\theta/x)+x^k(1-\mathrm{e}^{-\theta/x}),\quad \text{所有 } k,$$
$$\text{众数}=\theta/2,$$
$$\hat{\theta}=-q\ln(3/4).$$

A.4 其他概率分布

对数正态分布 —μ, σ(μ 可以取负值)

$$f(x)=\frac{1}{x\sigma\sqrt{2\pi}}\exp(-z^2/2)=\phi(z)/(\sigma x),\quad z=\frac{\ln x-\mu}{\sigma},$$
$$F(x)=\Phi(z),$$
$$\mathrm{E}[X^k]=\exp\left(k\mu+\frac{1}{2}k^2\sigma^2\right),$$
$$\mathrm{E}[(X\wedge x)^k]=\exp\left(k\mu+\frac{1}{2}k^2\sigma^2\right)\Phi\left(\frac{\ln x-\mu-k\sigma^2}{\sigma}\right)+x^k[1-F(x)],$$
$$\text{众数}=\exp(\mu-\sigma^2),$$
$$\hat{\sigma}=\sqrt{\ln(t)-2\ln(m)},\quad \hat{\mu}=\ln(m)-\frac{1}{2}\hat{\sigma}^2.$$

逆 Gaussian 分布—μ,θ

$$f(x)=\left(\frac{\theta}{2\pi x^3}\right)^{1/2}\exp\left(-\frac{\theta z^2}{2x}\right),\quad z=\frac{x-\mu}{\mu},$$
$$F(x)=\Phi\left[z\left(\frac{\theta}{x}\right)^{1/2}\right]+\exp\left(\frac{2\theta}{\mu}\right)\Phi\left[-y\left(\frac{\theta}{x}\right)^{1/2}\right],\quad y=\frac{x+\mu}{\mu},$$
$$\mathrm{E}[X]=\mu,\quad \mathrm{Var}[X]=\mu^3/\theta,$$
$$\mathrm{E}[X\wedge x]=x-\mu z\Phi\left[z\left(\frac{\theta}{x}\right)^{1/2}\right]-\mu y\exp(2\theta/\mu)\Phi\left[-y\left(\frac{\theta}{x}\right)^{1/2}\right],$$

$$M(t)=\exp\left[\frac{\theta}{\mu}\left(1-\sqrt{1-\frac{2\mu^2}{\theta}t}\right)\right],\quad t<\frac{\theta}{2\mu^2},$$
$$\hat{\mu}=m,\quad \hat{\theta}=\frac{m^3}{t-m^2}.$$

log-t 分布 —r, μ, σ (μ 可以取负值)

令 Y 为 r 个自由度的 t 分布, 则 $X=\exp(\sigma Y+\mu)$ 服从 $\log -t$ 分布. 这个分布不存在正数阶的矩, 正如 t 分布比正态分布尾部更厚一样, 这个分布比对数正态分布的尾部更厚.

$$f(x)=\frac{\Gamma(\frac{r+1}{2})}{x\sigma\sqrt{\pi r}\Gamma(\frac{r}{2})[1+\frac{1}{r}(\frac{\ln x-\mu}{\sigma})^2]^{(r+1)/2}},$$
$$F(x)=F_r\left(\frac{\ln x-\mu}{\sigma}\right)\ F_r(t)\text{是自由度为}\,r\,\text{的}\,t\,\text{分布密度函数的 cdf},$$

$$F(x)=\begin{cases}\dfrac{1}{2}\beta\left[\dfrac{r}{2},\dfrac{1}{2};\dfrac{r}{r+(\frac{\ln x-\mu}{\sigma})^2}\right], & 0<x\leqslant \mathrm{e}^{\mu},\\ 1-\dfrac{1}{2}\beta\left[\dfrac{r}{2},\dfrac{1}{2};\dfrac{r}{r+(\frac{\ln x-\mu}{\sigma})^2}\right], & x\geqslant \mathrm{e}^{\mu}.\end{cases}$$

单参数 Pareto 分布 —α, θ

$$f(x)=\frac{\alpha\theta^\alpha}{x^{\alpha+1}},\quad x>\theta,$$
$$F(x)=1-\left(\frac{\theta}{x}\right)^\alpha,\quad x>\theta,$$
$$\mathrm{E}[X^k]=\frac{\alpha\theta^k}{\alpha-k},\quad k<\alpha,$$
$$\mathrm{E}[(X\wedge x)^k]=\frac{\alpha\theta^k}{\alpha-k}-\frac{k\theta^\alpha}{(\alpha-k)x^{\alpha-k}},\quad x\geqslant\theta,$$
$$\text{众数}=\theta,$$
$$\hat{\alpha}=\frac{m}{m-\theta}.$$

注意: 尽管这里出现了两个参数, 但只有 α 是真正的参数, θ 的值必须预先确定.

A.5 有限支集分布

对如下的两个分布, 假设尺度参数 θ 已知.

广义 beta 分布 —$\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{\theta}$, $\boldsymbol{\tau}$

$$
\begin{aligned}
f(x) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} u^a (1-u)^{b-1} \frac{\tau}{x}, \quad 0 < x < \theta, \quad u = (x/\theta)^\tau, \\
F(x) &= \beta(a,b;u), \\
\mathrm{E}[X^k] &= \frac{\theta^k \Gamma(a+b)\Gamma(a+k/\tau)}{\Gamma(a)\Gamma(a+b+k/\tau)}, \quad k > -a\tau, \\
\mathrm{E}[(X \wedge x)^k] &= \frac{\theta^k \Gamma(a+b)\Gamma(a+k/\tau)}{\Gamma(a)\Gamma(a+b+k/\tau)} \beta(a+k/\tau, b; u) + x^k[1-\beta(a,b;u)].
\end{aligned}
$$

beta 分布 —$\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{\theta}$

$$
\begin{aligned}
f(x) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} u^a (1-u)^{b-1} \frac{1}{x}, \quad 0 < x < \theta, \quad u = x/\theta, \\
F(x) &= \beta(a,b;u), \\
\mathrm{E}[X^k] &= \frac{\theta^k \Gamma(a+b)\Gamma(a+k)}{\Gamma(a)\Gamma(a+b+k)}, \quad k > -a, \\
\mathrm{E}[X^k] &= \frac{\theta^k a(a+1)\cdots(a+k-1)}{(a+b)(a+b+1)\cdots(a+b+k-1)}, \quad k\text{为整数}, \\
\mathrm{E}[(X \wedge x)^k] &= \frac{\theta^k a(a+1)\cdots(a+k-1)}{(a+b)(a+b+1)\cdots(a+b+k-1)} \beta(a+k, b; u) \\
&\quad + x^k[1-\beta(a,b;u)], \\
\hat{a} &= \frac{\theta m^2 - mt}{\theta t - \theta m^2}, \quad \hat{b} = \frac{(\theta m - t)(\theta - m)}{\theta t - \theta m^2}.
\end{aligned}
$$

附录B 离散分布

B.1 引 言

这部分将 16 个离散分布归纳为 3 类, 这种分类是基于计算概率时所采用的算法. 对于很多比较熟悉的概率分布, 这些方程可能看上去与你所熟悉的不同, 但它们的确有相同的概率. 每个名称后都会给出参数, 除非特别说明, 所有参数都取正值. 在任何情况下, p_k 都表示观测到 k 次损失的概率.

最方便的形式是用阶乘矩表示各阶矩, j 阶阶乘矩 $\mu_{(j)} = \mathrm{E}[N(N-1)\cdots(N-j+1)]$, 由此得到 $\mathrm{E}[N] = \mu_{(1)}$, $\mathrm{Var}(N) = \mu_{(2)} + \mu_{(1)} - \mu_{(1)}^2$.

这里给出的估计量未必是有效的估计值, 而只是提供了一个最大化似然 (或其他) 函数的初值. 为了确定初值, 需要使用下面这些量 [n_k 表示观测值等于 k 的个数 (当 k 为最大的取值时, n_k 表示观测值大等于 k 的个数, 并假设这些观测值都恰好取 k), n 表示样本量大小]

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{\infty} k n_k, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{\infty} k^2 n_k - \hat{\mu}^2.$$

矩方法用上式确定初值时, 一般在参数上面加一个尖帽 (例如 $\hat{\lambda}$), 使用其他方法时, 则用带弯表示 (例如 $\tilde{\lambda}$). 当初值方程不能得到可接受的参数值时, 合理的简单推测是, 令所有 λ 和 β 参数的乘积等于样本均值, 而其他参数等于 1. 如果有两个 α 或 β 参数, 一个简单的选择是令每一个都等于样本均值的平方根.

最后一个需要介绍的概念是概率生成函数

$$P(z) = \mathrm{E}[z^N].$$

B.2 $(a, b, 0)$ 类分布

这一类概率分布的支集为 $0, 1, \cdots$. 对这类分布, 为确定一个特定的分布首先要具体给出 p_0, 然后利用 $p_k = (a + b/k)p_{k-1}$ 得到其他点的概率. 具体的分布可以通过 p_0, a, b 确定产生. 对任何分布有 $\mu_{(1)} = (a+b)/(1-a)$, 更高阶的 j 有 $\mu_{(j)} = (aj+b)\mu_{(j-1)}/(1-a)$. 方差为 $(a+b)/(1-a)^2$.

Poisson 分布 —$\boldsymbol{\lambda}$

$$
\begin{aligned}
p_0 &= \mathrm{e}^{-\lambda}, \quad a = 0, \quad b = \lambda, \\
p_k &= \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!}, \\
\mathrm{E}[N] &= \lambda, \quad \mathrm{Var}[N] = \lambda, \\
\hat{\lambda} &= \hat{\mu}, \\
P(z) &= \mathrm{e}^{\lambda(z-1)}.
\end{aligned}
$$

几何分布 —$\boldsymbol{\beta}$

$$
\begin{aligned}
p_0 &= \frac{1}{1+\beta}, \quad a = \frac{\beta}{1+\beta}, \quad b = 0, \\
p_k &= \frac{\beta^k}{(1+\beta)^{k+1}}, \\
\mathrm{E}[N] &= \beta, \quad \mathrm{Var}[N] = \beta(1+\beta), \\
\hat{\beta} &= \hat{\mu}, \\
P(z) &= [1-\beta(z-1)]^{-1}.
\end{aligned}
$$

这是负二项分布在 $r=1$ 时的特例.

二项分布 —$\boldsymbol{q, m}$, $(\mathbf{0 < q < 1}, \boldsymbol{m}$ 为整数)

$$
\begin{aligned}
p_0 &= (1-q)^m, \quad a = -\frac{q}{1-q}, \quad b = \frac{(m+1)q}{1-q}, \\
p_k &= \binom{m}{k} q^k (1-q)^{m-k}, \quad k = 0, 1, \cdots, m, \\
\mathrm{E}[N] &= mq, \quad \mathrm{Var}[N] = mq(1-q), \\
\hat{q} &= \hat{\mu}/m, \\
P(z) &= [1+q(z-1)]^m.
\end{aligned}
$$

负二项分布 —$\boldsymbol{\beta, r}$

$$
\begin{aligned}
p_0 &= (1+\beta)^{-r}, \quad a = \frac{\beta}{1+\beta}, \quad b = \frac{(r-1)\beta}{1+\beta}, \\
p_k &= \frac{r(r+1)\cdots(r+k-1)\beta^k}{k!(1+\beta)^{r+k}}, \\
\mathrm{E}[N] &= r\beta, \quad \mathrm{Var}[N] = r\beta(1+\beta), \\
\hat{\beta} &= \frac{\hat{\sigma}^2}{\hat{\mu}} - 1, \quad \hat{r} = \frac{\hat{\mu}^2}{\hat{\sigma}^2 - \hat{\mu}}, \\
P(z) &= [1-\beta(z-1)]^{-r}.
\end{aligned}
$$

B.3 $(a,b,1)$ 类分布

为了与 $(a,b,0)$ 类进行区分, $(a,b,1)$ 类的概率表示为 $\Pr(N=k)=p_k^M$ 或 $\Pr(N=k)=p_k^T$, 这取决于所在的子类. 在这个类中, p_0^M 是任意的 (也就是说, 它是一个参数), 而 p_1^M 或 p_1^T 是参数 a 和 b 的具体函数, 接下来的概率可以像 $(a,b,0)$ 类一样归纳得到 $p_k^M=(a+b/k)p_{k-1}^M$, $k=2,3,\cdots$, 对 p_k^T 也有相同的归纳计算. 这个分布类又可以分为下面的两个子类, 我们讨论这一类分布函数时, 总是考虑与之相对应的 $(a,b,0)$ 类分布, 它们有相同的 a 和 b. p_k 继续用来表示对应的 $(a,b,0)$ 类中的概率.

B.3.1 零点截断分布子类

这个类的成员满足 $p_0^T=0$, 所以不需要这个值, 只有变量不可能在零点取值时才使用这个概率分布. 一阶阶乘矩为 $\mu_{(1)}=(a+b)/[(1-a)(1-p_0)]$, 其中 p_0 是 $(a,b,0)$ 类对应分布的概率值. 对数分布 (没有对应的分布) 有 $\mu_{(1)}=\beta/\ln(1+\beta)$. 更高阶的阶乘矩可以使用 $(a,b,0)$ 类的方程归纳得到. 方差为 $(a+b)[1-(a+b+1)p_0]/[(1-a)(1-p_0)]^2$. 如果这个子类中的分布在 $(a,b,0)$ 类中有对应分布, 则 $p_k^T=p_k/(1-p_0)$.

零点截断 Poisson 分布—λ

$$
\begin{aligned}
p_1^T &= \frac{\lambda}{e^\lambda-1}, \quad a=0, \quad b=\lambda,\\
p_k^T &= \frac{\lambda^k}{k!(e^\lambda-1)},\\
\mathrm{E}[N] &= \lambda/(1-e^{-\lambda}), \quad \mathrm{Var}[N]=\lambda[1-(\lambda+1)e^{-\lambda}]/(1-e^{-\lambda})^2,\\
\tilde{\lambda} &= \ln(n\hat{\mu}/n_1),\\
P(z) &= \frac{e^{\lambda z}-1}{e^\lambda-1}.
\end{aligned}
$$

零点截断几何分布—β

$$
\begin{aligned}
p_1^T &= \frac{1}{1+\beta}, \quad a=\frac{\beta}{1+\beta}, \quad b=0,\\
p_k^T &= \frac{\beta^{k-1}}{(1+\beta)^k},\\
\mathrm{E}[N] &= 1+\beta, \quad \mathrm{Var}[N]=\beta(1+\beta),\\
\hat{\beta} &= \hat{\mu}-1,\\
P(z) &= \frac{[1-\beta(z-1)]^{-1}-(1+\beta)^{-1}}{1-(1+\beta)^{-1}}.
\end{aligned}
$$

这是零点截断负二项分布在 $r=1$ 时的特例.

对数分布—β

$$p_1^T=\frac{\beta}{(1+\beta)\ln(1+\beta)},\quad a=\frac{\beta}{1+\beta},\quad b=-\frac{\beta}{1+\beta},$$
$$p_k^T=\frac{\beta^k}{k(1+\beta)^k\ln(1+\beta)},$$
$$\mathrm{E}[N]=\beta/\ln(1+\beta),\quad \mathrm{Var}[N]=\frac{\beta[1+\beta-\beta/\ln(1+\beta)]}{\ln(1+\beta)},$$
$$\tilde{\beta}=\frac{n\hat{\mu}}{n_1}-1\quad\text{或}\quad\frac{2(\hat{\mu}-1)}{\hat{\mu}},$$
$$P(z)=1-\frac{\ln[1-\beta(z-1)]}{\ln(1+\beta)}.$$

这是零点截断负二项分布当 r 趋于 0 时的极限情况.

零点截断的二项分布 —$q,m(0<q<1,m$ 为整数)

$$p_1^T=\frac{m(1-q)^{m-1}q}{1-(1-q)^m},\quad a=-\frac{q}{1-q},\quad b=\frac{(m+1)q}{1-q},$$
$$p_k^T=\frac{\binom{m}{k}q^k(1-q)^{m-k}}{1-(1-q)^m},\quad k=1,2,\cdots,m,$$
$$\mathrm{E}[N]=\frac{mq}{1-(1-q)^m},$$
$$\mathrm{Var}[N]=\frac{mq[(1-q)-(1-q+mq)(1-q)^m]}{[1-(1-q)^m]^2},$$
$$\tilde{q}=\frac{\hat{\mu}}{m},$$
$$P(z)=\frac{[1+q(z-1)]^m-(1-q)^m}{1-(1-q)^m}.$$

零点截断的负二项分布 —$\beta,r,(r>-1,r\neq 0)$

$$p_1^T=\frac{r\beta}{(1+\beta)^{r+1}-(1+\beta)},\quad a=\frac{\beta}{1+\beta},\quad b=\frac{(r-1)\beta}{1+\beta},$$
$$p_k^T=\frac{r(r+1)\cdots(r+k-1)}{k![(1+\beta)^r-1]}\left(\frac{\beta}{1+\beta}\right)^k,$$
$$\mathrm{E}[N]=\frac{r\beta}{1-(1+\beta)^{-r}},$$
$$\mathrm{Var}[N]=\frac{r\beta[(1+\beta)-(1+\beta+r\beta)(1+\beta)^{-r}]}{[1-(1+\beta)^{-r}]^2},$$

$$\tilde{\beta} = \frac{\hat{\sigma^2}}{\hat{\mu}} - 1, \quad \tilde{r} = \frac{\hat{\mu}^2}{\hat{\sigma}^2 - \hat{\mu}},$$
$$P(z) = \frac{[1-\beta(z-1)]^{-r} - (1+\beta)^{-r}}{1-(1+\beta)^{-r}}.$$

这个概率分布有时也称为扩展截断负二项分布, 因为参数 r 被扩展到 0 以下.

B.3.2 零点修正分布子类

零点修正类分布首先为截断分布, 然后对零点任意设一个概率值, 其概率值 p_0^M 是一个参数, 其余概率再进行相应的调整. p_k^M 的值可由对应的零点截断概率分布得到: $p_k^M = (1-p_0^M)p_k^T$, 或者由对应的 $(a,b,0)$ 类分布得到: $p_k^M = (1-p_0^M)p_k/(1-p_0)$. 它也有与零点截断子类相同的递推关系式.

零点修正分布的均值是对应的零点截断分布均值的 $1-p_0^M$ 倍, 方差是零点截断方差的 $1-p_0^M$ 倍再加上零点截断均值平方的 $p_0^M(1-p_0^M)$ 倍. 概率生成函数为 $P^M(z) = p_0^M + (1-p_0^M)P(z)$, 其中 $P(z)$ 是对应的零点截断分布的概率生成函数.

一般情况下, p_0^M 的最大似然估计为样本在零点的相对频率.

B.4 复合分布类

这个分布类是通过一个概率分布与其他分布的复合得到的. 也就是说, 令 N 为一个离散分布, 称为**主分布**, 令 $M_1, M_2, \cdots$ 独立同分布, 分布函数为另一个离散分布, 称为**次分布**, 复合分布为 $S = M_1 + \cdots + M_N$ 的分布. 复合分布的概率可由下式得到:

$$p_k = \frac{1}{1-af_0}\sum_{y=1}^{k}(a+by/k)f_y p_{k-y},$$

$k = 1, 2, \cdots$, 其中 a 和 b 是主分布的值 [它必须是 $(a,b,0)$ 类], 而 f_y 是次分布的概率 p_y. 这里只采用两个主分布, Poisson 分布 $[p_0 = \exp[-\lambda(1-f_0)]]$ 和几何分布 $[p_0 = 1/(1+\beta-\beta f_0)]$. 因为名称本身可以完全确定分布, 所以下面只给出名称和初值.

复合分布的矩可由个体分布的矩得到:

$$\mathrm{E}[S] = \mathrm{E}[N]\mathrm{E}[M] \quad 和 \quad \mathrm{Var}[S] = \mathrm{E}[N]\mathrm{Var}[M] + \mathrm{Var}[N]\mathrm{E}[M]^2.$$

概率生成函数为 $P(z) = P_{主分布}[P_{次分布}(z)]$.

在下面列举分布的顺序是: 首先为主分布名称. 其中第一个、第二个和第四个概率分布的次分布使用 $(a,b,0)$ 类名称, 第三个和最后三个概率分布 (Poisson-ETNB 以及它的两个特例) 的次分布为零点截断的形式.

一些复合分布

Poisson- 二项分布 —$\boldsymbol{\lambda}$ 、$\boldsymbol{q}$、$\boldsymbol{m}$($\mathbf{0} < \boldsymbol{q} < \mathbf{1}, \boldsymbol{m}$为整数)

$$\hat{q} = \frac{\hat{\sigma}^2/\hat{\mu} - 1}{m - 1}, \quad \hat{\lambda} = \frac{\hat{\mu}}{m\hat{q}} \quad \text{或} \quad \tilde{q} = 0.5, \quad \tilde{\lambda} = \frac{2\hat{\mu}}{m}.$$

Poisson-Poisson 分布—$\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$

λ_1 是主 Poisson 分布的参数, λ_2 是次 Poisson 分布的参数, 也称这个分布为 **Neyman Type A**.

$$\tilde{\lambda}_1 = \tilde{\lambda}_2 = \sqrt{\hat{\mu}}.$$

几何–扩展截尾负二项分布 —$\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2, \boldsymbol{r}(\boldsymbol{r} > -\mathbf{1})$

β_1 是主几何分布的参数, 后两个是次分布的参数, 注意 $r = 0$ 时次分布为对数分布. 由于使用截断形式, 所以 r 的扩展是有效的.

$$\tilde{\beta}_1 = \tilde{\beta}_2 = \sqrt{\hat{\mu}}.$$

几何 -Poisson 分布 —$\boldsymbol{\beta}$, $\boldsymbol{\lambda}$

$$\tilde{\beta} = \tilde{\lambda} = \sqrt{\hat{\mu}}.$$

Poisson- 扩展截尾负二项分布 — $\boldsymbol{\lambda}, \boldsymbol{\beta}, \boldsymbol{r}, (\boldsymbol{r} > -1, \boldsymbol{r} \neq 0)$

当 $r = 0$ 时次分布为对数分布, 复合得到负二项分布.

$$\tilde{r} = \frac{\hat{\mu}(K - 3\hat{\sigma}^2 + 2\hat{\mu}) - 2(\hat{\sigma}^2 - \hat{\mu})^2}{\hat{\mu}(K - 3\hat{\sigma}^2 + 2\hat{\mu}) - (\hat{\sigma}^2 - \hat{\mu})^2}, \quad \tilde{\beta} = \frac{\hat{\sigma}^2 - \hat{\mu}}{\hat{\mu}(1 + \hat{r})}, \quad \tilde{\lambda} = \frac{\hat{\mu}}{\hat{r}\hat{\beta}},$$

或

$$\tilde{r} = \frac{\hat{\sigma}^2 n_1/n - \hat{\mu}^2 n_0/n}{(\hat{\sigma}^2 - \hat{\mu}^2)(n_0/n)\ln(n_0/n) - \hat{\mu}(\hat{\mu}n_0/n - n_1/n)},$$

$$\tilde{\beta} = \frac{\hat{\sigma}^2 - \hat{\mu}}{\hat{\mu}(1 + \hat{r})}, \quad \tilde{\lambda} = \frac{\hat{\mu}}{\hat{r}\hat{\beta}},$$

其中

$$K = \frac{1}{n}\sum_{k=0}^{\infty} k^3 n_k - 3\hat{\mu}\frac{1}{n}\sum_{k=0}^{\infty} k^2 n_k + 2\hat{\mu}^3.$$

也称这个分布为**广义 Poisson- 帕斯卡分布**.

Polya-Aeppli 分布—$\boldsymbol{\lambda}$, $\boldsymbol{\beta}$

$$\hat{\beta} = \frac{\hat{\sigma}^2 - \hat{\mu}}{2\hat{\mu}}, \quad \hat{\lambda} = \frac{\hat{\mu}}{1 + \hat{\beta}}.$$

这是 Poisson 与扩展截断的负二项分布复合时在 $r = 1$ 时的特例, 实际上它是 Poisson 与截断的几何分布的复合.

Poisson- 逆高斯分布 —λ, β

$$\tilde{\lambda} = -\ln(n_0/n), \quad \tilde{\beta} = \frac{4(\hat{\mu} - \hat{\lambda})}{\hat{\mu}}.$$

这是 Poisson 与扩展截断负二项分布复合时, $r = -0.5$ 时的特例.

B.5 离散分布的汇总

下面的表格显示有些概率分布是其他分布的特例或极限情况. 所谓特例是将一个参数设为常数, 极限情况是将两个参数按某些方式趋于无穷或零.

概率分布	分布的特殊情况	分布的极限情况
Poisson	ZM Poisson	负二项, Poisson- 二项, Poisson-inv. Gaussian, Polya-Aeppli, Neyman-A
ZT Poisson	ZM Poisson	ZT 负二项
ZM Poisson		ZM 负二项
几何	负二项 ZM 几何	几何 -Poisson
ZT 几何	ZT 负二项	
ZM 几何	ZM 负二项	
对数		ZT 负二项
ZM 对数		ZM 负二项
二项	ZM 二项	
负二项	ZM 负二项	Poisson-ETNB
Poisson-inverse Gaussian	Poisson-ETNB	
Polya-Aeppli	Poisson-ETNB	
Neyman-A		Poisson-ETNB

附录C　损失频率和损失程度的关系

令 N^L 表示损失次数随机变量, X 表示损失量随机变量. 如果引入免赔额 d, 有两种方法对 X 进行修正. 一种方法是构造 Y^L, 每次损失的赔付额表示为

$$Y^L=\begin{cases}0, & X\leqslant d,\\ X-d, & X>d.\end{cases}$$

这种情况下, 相应的频率变量仍然为 N^L.

另一种方法是建立 Y^P, 表示每次赔付的赔付总量为

$$Y^P=\begin{cases}\text{未定义}, & X\leqslant d,\\ X-d, & X>d.\end{cases}$$

这时必须改变相应的频率变量以反映赔付数, 设这个变量为 N^P. 首先假设每次损失引起赔付的概率为 $v=1-F_X(d)$, 再假设赔付与损失数是相互独立的. 则 $N^P=L_1+L_2+\cdots+L_N$, 其中 L_j 以概率 $1-v$ 等于 0, 以概率 v 等于 1. 通过概率生成函数可得到如下关系:

N^L 的分布	N^P 的参数
Poisson	$\lambda^*=v\lambda$
ZM Poisson	$p_0^{M*}=\dfrac{p_0^M-\mathrm{e}^{-\lambda}+\mathrm{e}^{-v\lambda}-p_0^M\mathrm{e}^{-v\lambda}}{1-\mathrm{e}^{-\lambda}}, \lambda^*=v\lambda$
二项	$q^*=vq$
ZM 二项	$p_0^{M*}=\dfrac{p_0^M-(1-q)^m+(1-vq)^m-p_0^M(1-vq)^m}{1-(1-q)^m}$
	$q^*=vq$
负二项	$\beta^*=v\beta, r^*=r$
ZM 负二项	$p_0^{M*}=\dfrac{p_0^M-(1+\beta)^{-r}+(1+v\beta)^{-r}-p_0^M(1+v\beta)^{-r}}{1-(1+\beta)^{-r}}$
	$\beta^*=v\beta, r^*=r$
ZM 对数	$p_0^{M*}=1-(1-p_0^M)\ln(1+v\beta)/\ln(1+\beta)$
	$\beta^*=v\beta$

这里没有列举几何分布, 因为它是负二项分布在 $r=1$ 时的特例. 对于零点截断概率分布, 上表仍然用变量 N^P 表示, 它为零点修正分布. 复合分布中只修正了次分布, 次分布为 ETNB 分布时, 主分布的参数要乘以 $1-p_0^{M*}$, 使得次分布仍保持为零点截断的形式 $(\beta^*=v\beta)$.

有些情况下收集到的频率数据只能形成 N^P 的模型, 这时为了得到有效的 v 就不得不确定一个免赔额 d. 重新得到 N^L 的概率分布是可能的, 尽管并不保证反过来的过程会产生一个合理的概率分布. 解法和上面相同, 只是现在的 $v=1/[1-F_X(d)]$.

假设当期的频率模型为 N^d, 与免赔额 d 相对应. 若免赔额变为 d^*, 则赔付频率为 N^{d^*}, 它与 N^d 具有相同的形式, 根据上表知 $v=[1-F_X(d^*)]/[1-F_X(d)]$.

附录D　递 归 公 式

递归公式为 (其中的索赔频率分布属于 $(a, b, 1)$ 类):

$$f_S(x) = \frac{[p_1 - (a+b)p_0]f_X(x) + \sum_{y=1}^{x \wedge m}(a + \frac{by}{x})f_X(y)f_S(x-y)}{1 - af_X(0)},$$

其中, $f_S(x) = \Pr(S = x)$, $x = 0, 1, 2, \cdots$, $f_X(x) = \Pr(X = x)$, $x = 0, 1, 2, \cdots$, $p_0 = \Pr(N = 0)$ 和 $p_1 = \Pr(N = 1)$. 注意损失程度变量 (X) 的概率必须定义于非负整数, 公式的初值为 $f_S(0)$, 这些值在表 D-1 中给出. 应当注意到, 若 N 属于 $(a, b, 0)$ 类, $p_1 - (a+b)p_0 = 0$, 所以第一项为 0. 若 N 是复合类分布, 须运行递归公式两次. 第一次利用 p_0, p_1, a, b 对次分布递归得到 $f_X(x)$, 第二次通过 p_0, p_1, a, b 对主分布递归.

表 D-1　递归公式的初始值 ($f_S(0)$)

分　布	$f_S(0)$
Poisson	$\exp[\lambda(f_0 - 1)]$
几何	$[1 + \beta(1 - f_0)]^{-1}$
二项分布	$[1 + q(f_0 - 1)]^m$
负二项分布	$[1 + \beta(1 - f_0)]^{-r}$
ZM Poisson	$p_0^M + (1 - p_0^M)\dfrac{\exp(\lambda f_0) - 1}{\exp(\lambda) - 1}$
ZM 几何	$p_0^M + (1 - p_0^M)\dfrac{f_0}{1 + \beta(1 - f_0)}$
ZM 二项	$p_0^M + (1 - p_0^M)\dfrac{[1 + q(f_0 - 1)]^m - (1 - q)^m}{1 - (1 - q)^m}$
ZM 负二项	$p_0^M + (1 - p_0^M)\dfrac{[1 + \beta(1 - f_0)]^{-r} - (1 + \beta)^{-r}}{1 - (1 + \beta)^{-r}}$
ZM 对数	$p_0^M + (1 - p_0^M)\left\{1 - \dfrac{\ln[1 + \beta(1 - f_0)]}{\ln(1 + \beta)}\right\}$

附录E　损失程度分布的离散化方法

有两种相对简单的方法对损失程度分布进行离散化处理, 一种是取整法, 另一种是均值不变法.

E.1　取　整　法

这种方法有两个特点: 所有的概率都是正的, 所有的概率和为 1. 令 h 为跨度, Y 为 X 离散化后的形式. 若没有任何调整, 则

$$\begin{aligned} f_j = \Pr(Y = jh) &= \Pr\left[\left(j - \frac{1}{2}\right)h \leqslant X < \left(j + \frac{1}{2}\right)h\right] \\ &= F_X\left[\left(j + \frac{1}{2}\right)h\right] - F_X\left[\left(j - \frac{1}{2}\right)h\right]. \end{aligned}$$

然后对 $f_X(j) = f_j$ 使用递归公式. 假设免赔额为 d, 保单限额为 u, 并且共同保险比例为 α, 若在离散化前进行这些调整, 则

$$\begin{aligned} g_0 &= \frac{F_X(d + h/2) - F_X(d)}{1 - F_X(d)}, \\ g_j &= \frac{F_X[d + (j + 1/2)h] - F_X[d + (j - 1/2)h]}{1 - F_X(d)}, \\ &\quad j = 1, \cdots, \frac{u - d}{h} - 1, \\ g_{(u-d)/h} &= \frac{1 - F_X(u - h/2)}{1 - F_X(d)}, \end{aligned}$$

其中, $g_j = \Pr(Z = j\alpha h)$, Z 是调整后的分布. 这种方法不需要保单限额是 h 的整数倍, 但要求 $u - d$ 是 h 的整倍数. 这种方法给出了每单赔案所有可能赔付额的概率.

最后, 若在 u 点以上截断, 则将所有的分母改为 $F_X(u) - F_X(d)$, 并将 $g_{(u-d)/h}$ 的分子改为 $F_X(u) - F_X(u - h/2)$.

E.2　均值不变法

这种方法使得离散化后的分布与原始损失程度分布有相同的均值, 如果没有调整, 离散化形式为

$$f_0 = 1 - \frac{\mathrm{E}[X \wedge h]}{h},$$
$$f_j = \frac{2\mathrm{E}[X \wedge jh] - \mathrm{E}[X \wedge (j-1)h] - \mathrm{E}[X \wedge (j+1)h]}{h}, \quad j = 1, 2, \cdots.$$

调整后的分布有

$$g_0 = 1 - \frac{\mathrm{E}[X \wedge d + h] - \mathrm{E}[X \wedge d]}{h[1 - F_X(d)]},$$
$$g_j = \frac{2\mathrm{E}[X \wedge d + jh] - \mathrm{E}[X \wedge d + (j-1)h] - \mathrm{E}[X \wedge d + (j+1)h]}{h[1 - F_X(d)]},$$
$$j = 1, \cdots, \frac{u-d}{h} - 1,$$
$$g_{(u-d)/h} = \frac{\mathrm{E}[X \wedge u] - \mathrm{E}[X \wedge u - h]}{h[1 - F_X(d)]}.$$

为了适合右尾部的截断, 分母改为

$$h[F_X(u) - F_X(d)],$$

并从分子中的每个 g_0 和 $g_{(u-d)/h}$ 中减去 $h[1 - F_X(u)]$.

E.3 离散分布的连续化

假设已有 $g_0 = \Pr(S = 0)$, 即随机变量等于 0 的真实概率. 令 $p_j = \Pr(S^* = jh)$, 其中 S^* 为离散的, h 为跨度. 下面对 S(真实分布, 离散化后的形式为 S^*) 近似累积分布函数和 LEV. 假设对每个区间 j, S 在 $(j - \frac{1}{2})h$ 到 $(j + \frac{1}{2})h$ 的区间上为均匀分布. 第一个区间从 0 到 $h/2$, 假设概率 $p_0 - g_0$ 均匀地分布其中. 记 S^{**} 是具有这种近似混合分布的随机变量 (连续的, 除 0 点有离散概率 g_0). 可以通过如下的插值法来建立近似分布函数, 首先, 令

$$F_j = F_{S^{**}}\left[\left(j + \frac{1}{2}\right)h\right] = \sum_{i=0}^{j} p_i, \quad j = 0, 1, \cdots.$$

然后, 设 x 在区间 $(j - \frac{1}{2})h$ 到 $(j + \frac{1}{2})h$ 中

$$\begin{aligned}
F_{S^{**}}(x) &= F_{j-1} + \int_{(j-1/2)h}^{x} h^{-1} p_j \mathrm{d}t = F_{j-1} + \left[x - \left(j - \frac{1}{2}\right)h\right] h^{-1} p_j \\
&= F_{j-1} + \left[x - \left(j - \frac{1}{2}\right)h\right] h^{-1}(F_j - F_{j-1}) \\
&= (1 - w)F_{j-1} + wF_j, \quad w = \frac{x}{h} - j + \frac{1}{2}.
\end{aligned}$$

由于第一个区间宽度只有一半, 所以对 $0 \leqslant x \leqslant h/2$, 公式为

$$F_{S^{**}}(x) = (1-w)g_0 + wp_0, \quad w = \frac{2x}{h}.$$

也可以用离散概率形式表示这些公式

$$F_{S^{**}}(x) = \begin{cases} g_0 + \dfrac{2x}{h}[p_0 - g_0], & 0 < x \leqslant \dfrac{h}{2}, \\ \displaystyle\sum_{i=0}^{j-1} p_i + \frac{x-(j-1/2)h}{h} p_j, & \left(j-\dfrac{1}{2}\right)h < x \leqslant \left(j+\dfrac{1}{2}\right)h. \end{cases}$$

考虑含限额的期望值 (LEV), 一阶和 k 阶 LEV 的表达式为

$$\mathrm{E}(S^{**} \wedge x) = \begin{cases} x(1-g_0) - \dfrac{x^2}{h}(p_0 - g_0), & 0 < x \leqslant \dfrac{h}{2}, \\ \dfrac{h}{4}(p_0 - g_0) + \displaystyle\sum_{i=1}^{j-1} ihp_i + \frac{x^2 - [(j-1/2)h]^2}{2h} p_j, & \\ \quad + x[1 - F_{S^{**}}(x)], & \left(j-\dfrac{1}{2}\right)h < x \leqslant \left(j+\dfrac{1}{2}\right)h. \end{cases}$$

对 $0 < x \leqslant \frac{h}{2}$

$$\mathrm{E}[(S^{**} \wedge x)^k] = \frac{2x^{k+1}}{h(k+1)}(p_0 - g_0) + x^k[1 - F_{S^{**}}(x)],$$

对 $(j-\frac{1}{2})h < x \leqslant (j+\frac{1}{2})h$

$$\begin{aligned} \mathrm{E}[(S^{**} \wedge x)^k] = {} & \frac{(h/2)^k(p_0 - g_0)}{k+1} + \sum_{i=1}^{j-1} \frac{h^k[(i+\frac{1}{2})^{k+1} - (i-\frac{1}{2})^{k+1}]}{k+1} p_i \\ & + \frac{x^{k+1} - [(j-\frac{1}{2})h]^{k+1}}{h(k+1)} p_j + x^k[1 - F_{S^{**}}(x)]. \end{aligned}$$

附录F　数值优化和方程组求解

对包含很多变量的函数进行最大化求解通常是非常困难的, 目前已经产生了许多的数值方法, 其中的大多数都可以解决本书提出的问题. 这里提供了两种方法. 第一个是嵌在 Excel® 中的 Solver 程序 (简称**解算器**). 这个方法一般情况下相当可靠, 尽管有时会对没有最大值的情形给出一个最大值. 第二个选择是单纯型方法. 这个方法相对比较慢, 但是更加可靠. 在本附录的最后一节将演示如何使用 Excel® Solver 程序求解方程组.

F.1　使用 Solver 程序求解最大化

安装了 Excel® 后并不能自动实现解算器的运算. 如果可用则可以在 Excel 的工具菜单中找到. 如果没有安装, 要在工具菜单中选择加载宏, 然后选中解算器求解复选框, 点击确定. 如果解算器求解没有出现在可用加载宏的列表中, 说明安装 Excel® 的时候没有安装解算器程序, 一般使用典型 (不是完全或自定义) 安装将会产生这样的后果. 为了安装解算器求解, 选择控制面板中的添加 (删除) 程序, 然后修改 Microsoft Office® 的安装. 不需要为了加载解算器而重新安装 Office®.

Microsoft Excel - solver illustration.xls

File　Edit　View　Insert　Format　Tools　Data　Window　Help

E3　=SUM(C2:C8)

	A	B	C	D	E	F
1	x	"f(x)"	ln	alpha	1	
2	27	0.000973	-6.93476	theta	1000	
3	82	0.000921	-6.98976	lnL	-44.9125	
4	115	0.000891	-7.02276			
5	126	0.000882	-7.03376			
6	155	0.000856	-7.06276			
7	161	0.000851	-7.06876			
8	200	0.818731	-2.8			
9						
10						
11						
12						
13						

Sheet1　Sheet2　Sheet3

Ready

下面将通过一个例子来演示解算器求解的使用：使用数据集 B 且右删失点为 200 的数据，求解 gamma 模型的最大似然估计. 如果读者并不很理解这个例子，也并不重要，这里只是为了说明.

首先在一个电子表格中设置目标函数 (lnL) 的参数 (alpha 和 theta) 的单元格. 本例中这两个参数分别位于 E1 和 E2 中，而目标函数在 E3 中.[①]

这个电子表格的公式显式如下

Microsoft Excel - solver illustration.xls

File Edit View Insert Format Tools Data Window Help

E7 fx

	A	B	C	D	E
1	x	"f(x)"	ln	alpha	1
2	27	=GAMMADIST(A2,E$1,E$2,FALSE)	=LN(B2)	theta	1000
3	82	=GAMMADIST(A3,E$1,E$2,FALSE)	=LN(B3)	lnL	=SUM(C2:C8)
4	115	=GAMMADIST(A4,E$1,E$2,FALSE)	=LN(B4)		
5	126	=GAMMADIST(A5,E$1,E$2,FALSE)	=LN(B5)		
6	155	=GAMMADIST(A6,E$1,E$2,FALSE)	=LN(B6)		
7	161	=GAMMADIST(A7,E$1,E$2,FALSE)	=LN(B7)		
8	200	=1-GAMMADIST(200,E1,E2,TRUE)	=14*LN(B8)		
9					
10					

Sheet1 / Sheet2 / Sheet3

Ready

注意 alpha 和 beta 的初值已经输入 (1 和 1 000). 这些初值越合适，解算器找到最大值的可能性就越大. 由工具菜单选择解算器求解将打开如下对话框：

Solver Parameters

Set Target Cell: E3

Equal To: Max Min Value of: 0

By Changing Cells:

E1,E2

Subject to the Constraints:

Solve

Close

Guess

Options

Add

Change

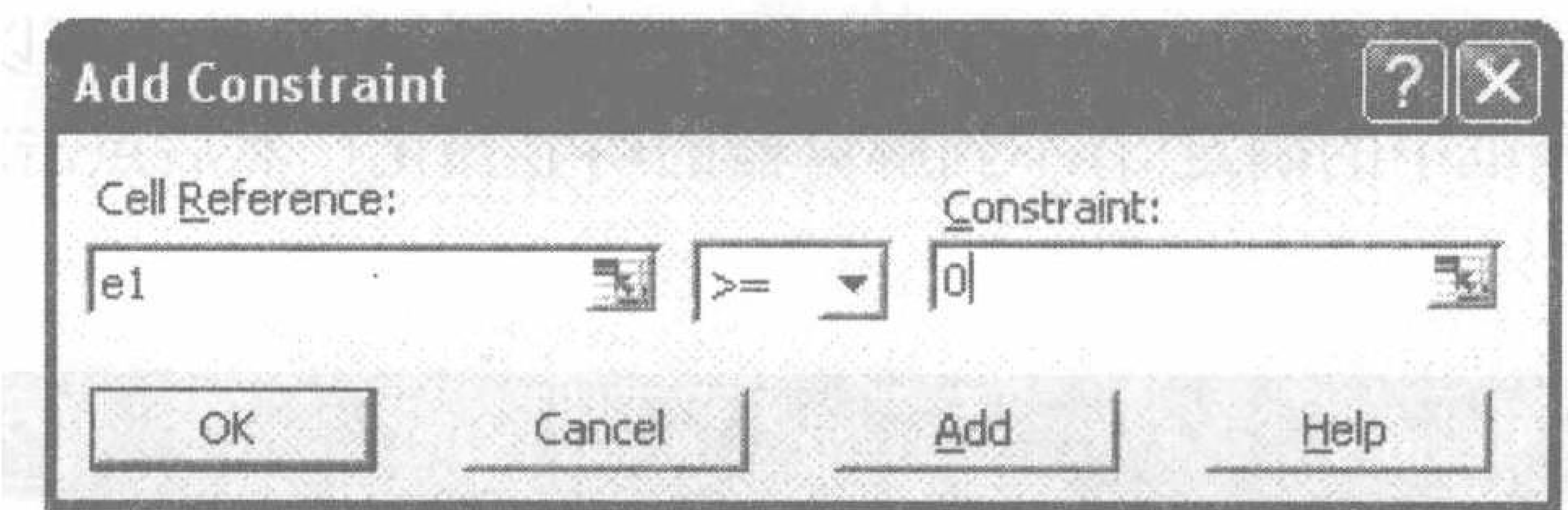
Delete

Reset All

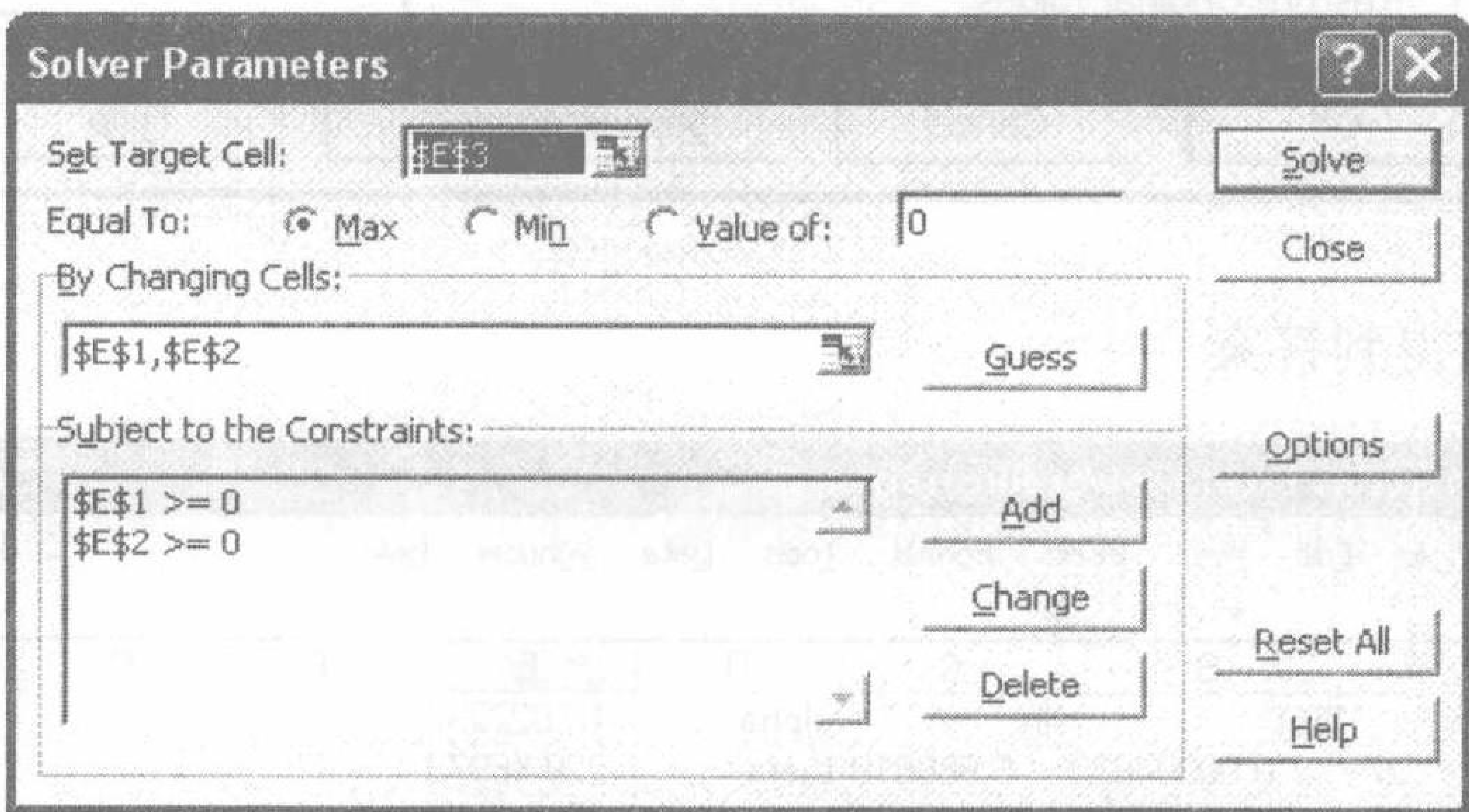
Help

目标单元格为目标函数所在的位置，**可变单元格**包含了可变参数的区域，并不需要这些单元格是相邻的. 点击**求解**就可以得到答案，但是还要注意下面两个问题. 第一，解算器求解允许带约束条件. 可以通过的单击**添加**得到如下的对话框：

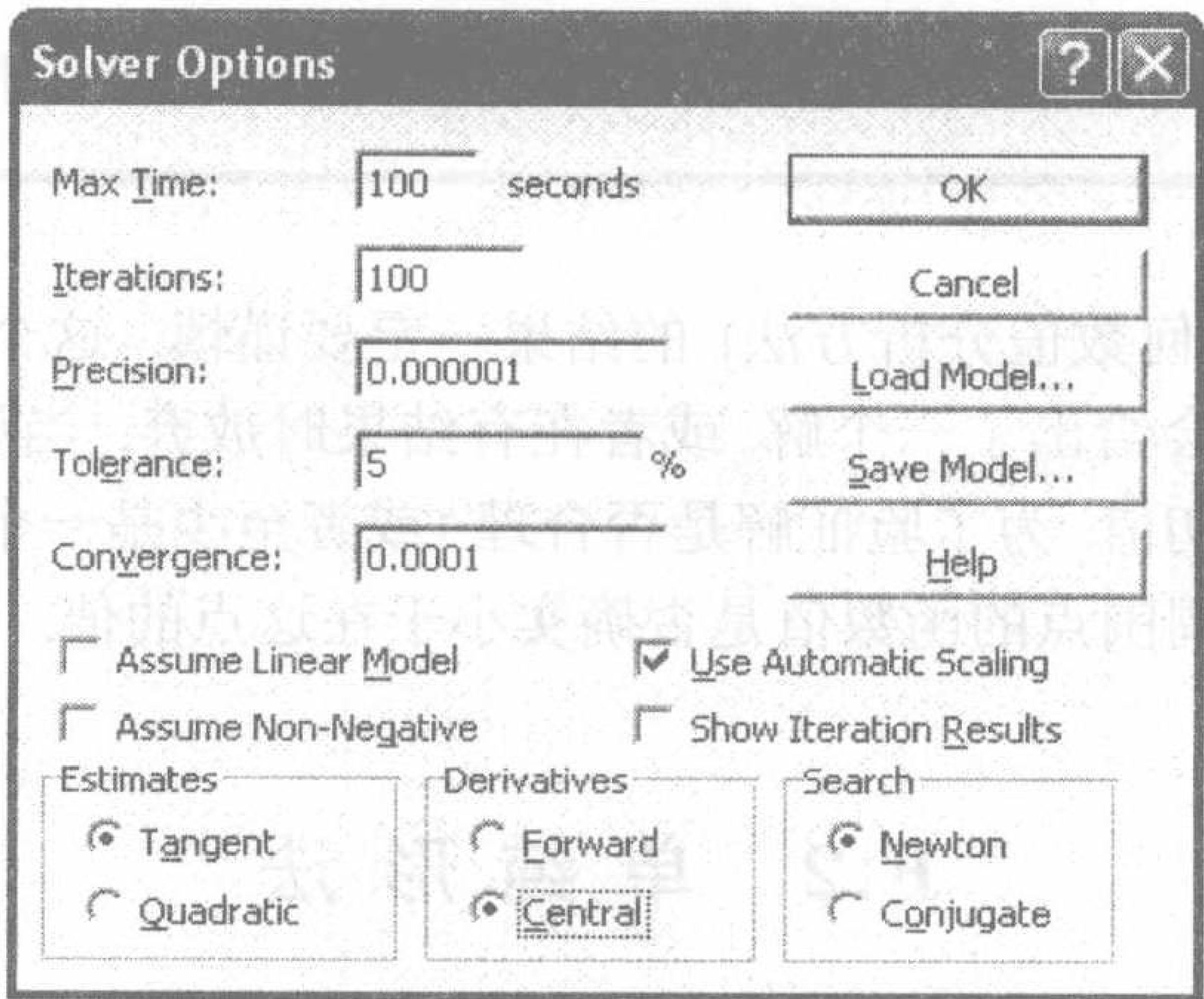
① 屏幕截取的图像已得到微软公司的版权许可.



这里添加了约束条件 $\alpha \geqslant 0$. 解算器并不能够执行我们真正想要的约束 ($\alpha > 0$). 类似地, 在输入了 θ 的约束后, 解算器求解的对话框显示如下:

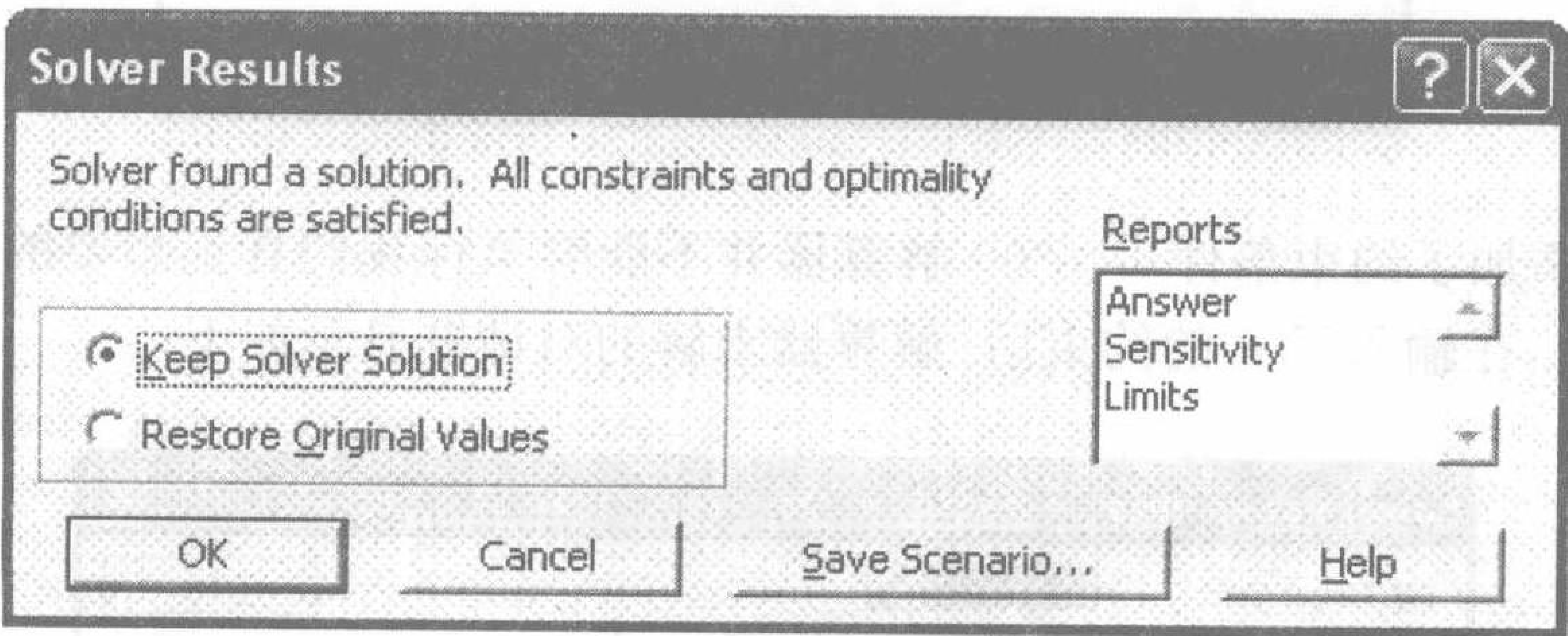


这里无需说明添加这些约束的原因, 只要解算器找到的值是满足约束条件的. 点击**选项**将出现下面的对话框:

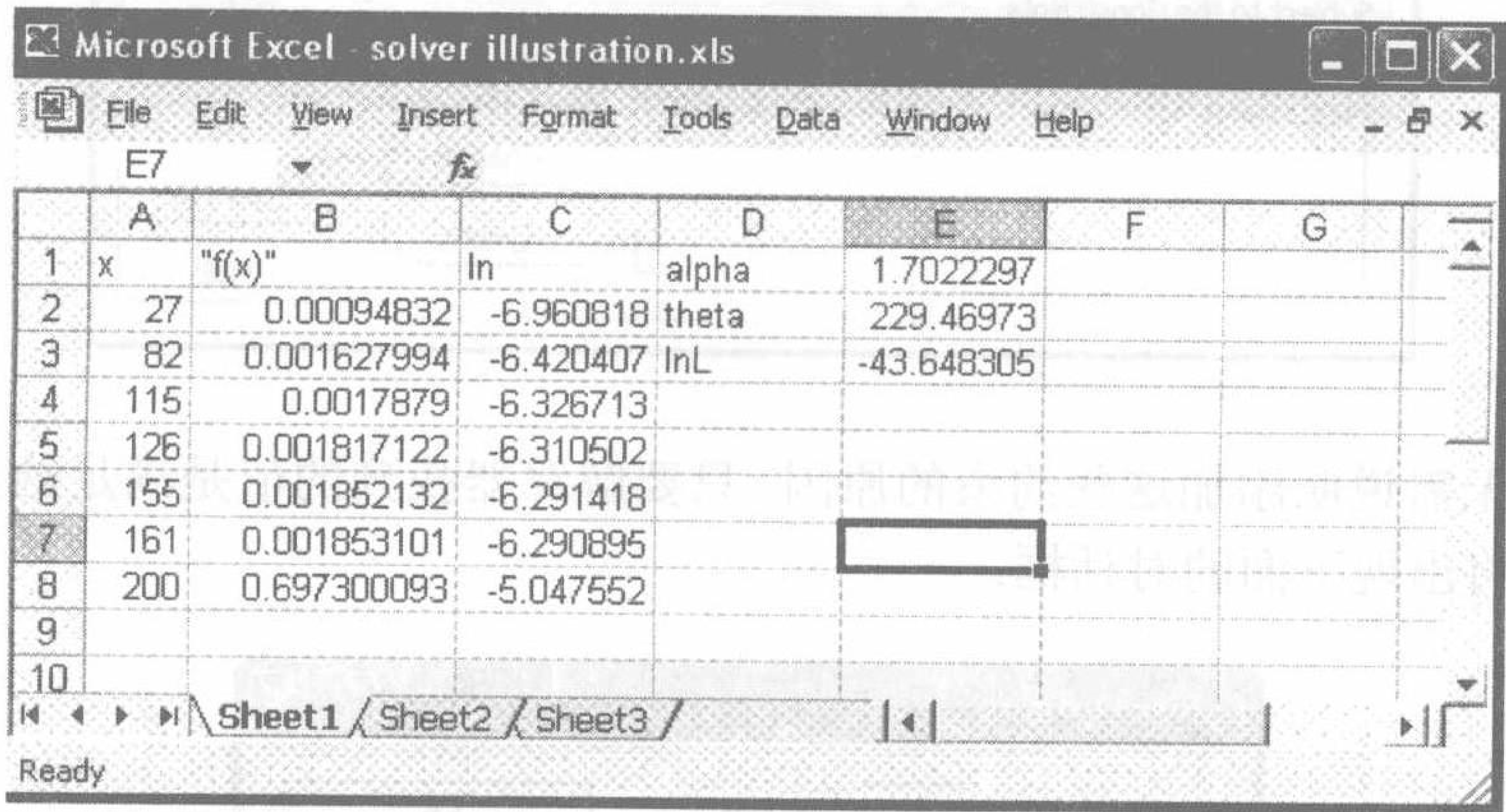


对于默认设置上述选项修改了两个地方. 其一为选择**自动按比例缩放**, 当参数在尺度上无差异时 (如这个例子的情况) 这个选项可以提高性能. 另外的一个选择

是**中心化**导数近似, 通过选择更小的精度值、允许误差和收敛数可以得到更高精度的结果. 点击选择框中的**确定**(看不到解算器的明显变化), 然后点击**求解**, 将出现如下对话框:



点击**确定**得到答案



	A	B	C	D	E	F	G
1	x	"f(x)"	ln	alpha	1.7022297		
2	27	0.00094832	-6.960818	theta	229.46973		
3	82	0.001627994	-6.420407	lnL	-43.648305		
4	115	0.0017879	-6.326713				
5	126	0.001817122	-6.310502				
6	155	0.001852132	-6.291418				
7	161	0.001853101	-6.290895				
8	200	0.697300093	-5.047552				
9							
10							

对**解算器**(或任何数值分析方法) 的结果一定要谨慎. 这个程序可能对没有找到最大值的情况也会给出了一个解, 或者在有结果时放弃. 当程序放弃的时候, 可能需要一个更优的初值. 为了验证解是否合理 (或者至少是一个局部最大值), 一个有效的方法是检查周围点的函数值是否确实小于在这点的值.

F.2　单 纯 形 法

这个方法 (和运筹中的单纯型法不一样) 在 1965 年由 Nelder and Mead[98] 引入求解最大似然估计. Walters, Parker, Morgan 和 Deming 的专著*Sequential Simplex Optimization* [134] 是一本很好的参考书 (也是这里讨论的来源).

令 $\boldsymbol{x}$ 为 $k\times 1$ 向量, $f(\boldsymbol{x})$ 为待求的函数. 迭代过程开始于 $k+1$ 个向量: $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{k+1}$, 对应的函数值为 $f_1,\cdots,f_{k+1}$. 每一步迭代时将这些点排序, 使得 $f_2<\cdots<f_{k+1}$, 最初, 令 $f_1<f_2$. 并对三个点命名, $\boldsymbol{x}_1$ 称为**最差点**, $\boldsymbol{x}_2$ 称为**第二差点**, $\boldsymbol{x}_{k+1}$ 称为**最优点**. 需要注意的是经过了第一次的迭代, 这些名字可能并不能很好地描述这些点. 现在确定 5 个新点. 第一个 y_1 是 $x_2,\cdots,x_{k+1}$ 的中心, 即 $y_1=\sum_{j=2}^{k+1}x_j/k$, 称为**中点**. 另外 4 个点如下得到

$$\boldsymbol{y}_2=2\boldsymbol{y}_1-\boldsymbol{x}_1,\ \textbf{参考点},$$

$$\boldsymbol{y}_3=2\boldsymbol{y}_2-\boldsymbol{x}_1,\ \textbf{双重点},$$

$$\boldsymbol{y}_4=(\boldsymbol{y}_1+\boldsymbol{y}_2)/2,\ \textbf{半点},$$

$$\boldsymbol{y}_5=(\boldsymbol{y}_1+\boldsymbol{x}_1)/2,\ \textbf{中心点}.$$

然后令 $g_2,\cdots,g_5$ 为对应的函数值, 即 $g_j=f(y_j)$(y_1 的值从未用过). 下一步的关键是选择这些点中的一个代替最差点. 决策过程如下.

(1) 如果 $f_2<g_2<f_{k+1}$, 则用**参考点**替换最差点.

(2) 如果 $g_2\geqslant f_{k+1}$ 且 $g_3>f_{k+1}$, 则用**双重点**替换最差点.

(3) 如果 $g_2\geqslant f_{k+1}$ 且 $g_3\leqslant f_{k+1}$, 则用**参考点**替换最差点.

(4) 如果 $f_1<g_2\leqslant f_2$, 则用**半点**替换最差点.

(5) 如果 $g_2\leqslant f_1$, 则用**中心点**替换最差点.

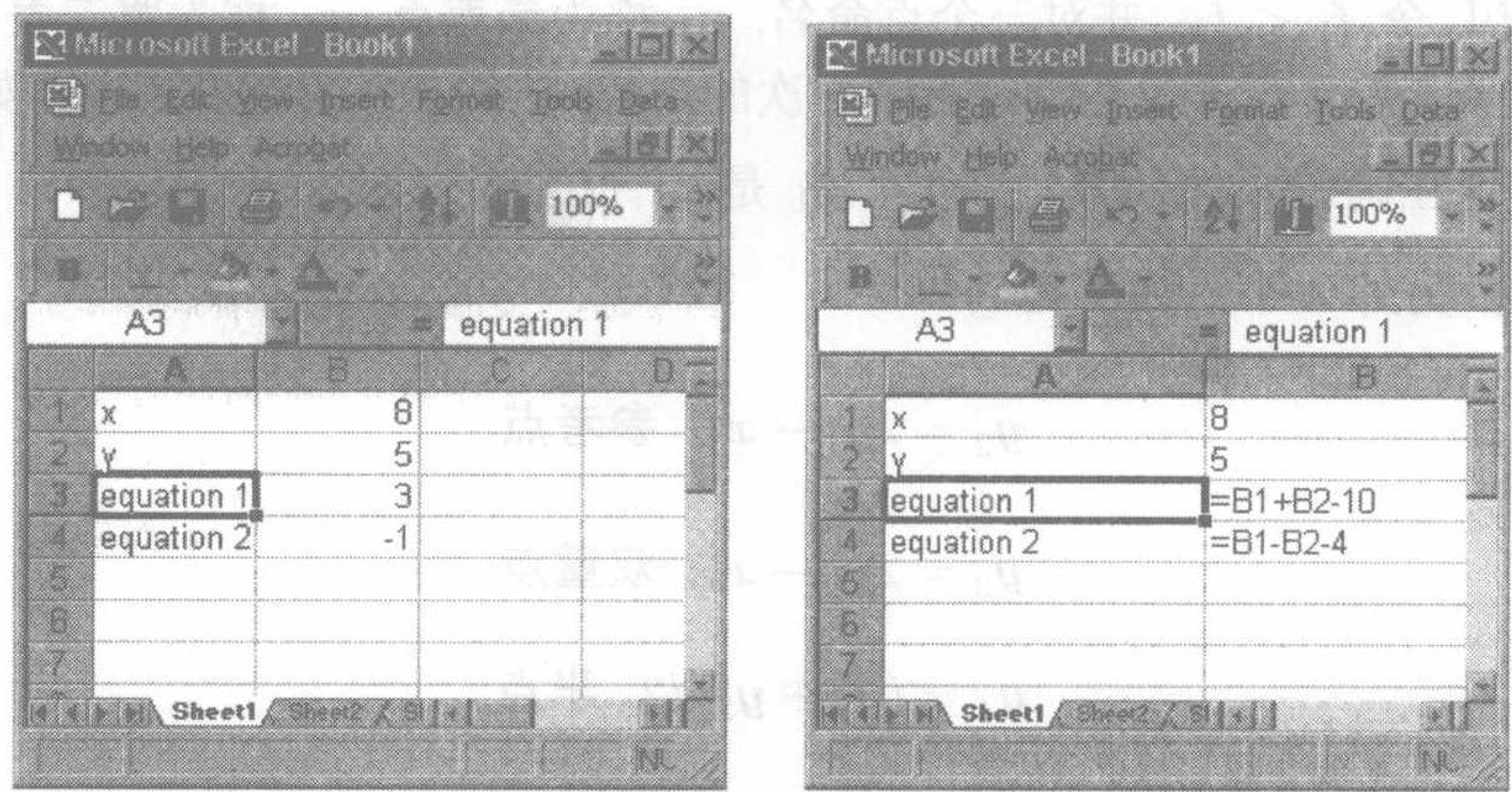
完成了替换的操作后, 原来的**第二差点**变成了**最差点**. 将剩下的 k 个点排序, 具有最小函数值的点称为新的**第二差点**, 最大函数值的点变成新的**最佳点**. 在操作中, 当达到第二步的时候才有必要计算 $\boldsymbol{y}_3$ 和 g_3. 还要注意在 $(\boldsymbol{y}_4,g_4)$ 和 $(\boldsymbol{y}_5,g_5)$ 这两对中至多有一个需要计算, 这依赖于第 4 步还是第 5 步的条件成立.

不断进行迭代, 直到 $k+1$ 个点的集合变得很紧, 有很多方法可以度量紧密的程度. 一个方法是计算每个分量的标准差, 然后取平均值. 当达到足够小的值时迭代停止. 另一个方法是不断进行迭代, 直到所有的 $k+1$ 个向量都达到了一个确定的有效数字.

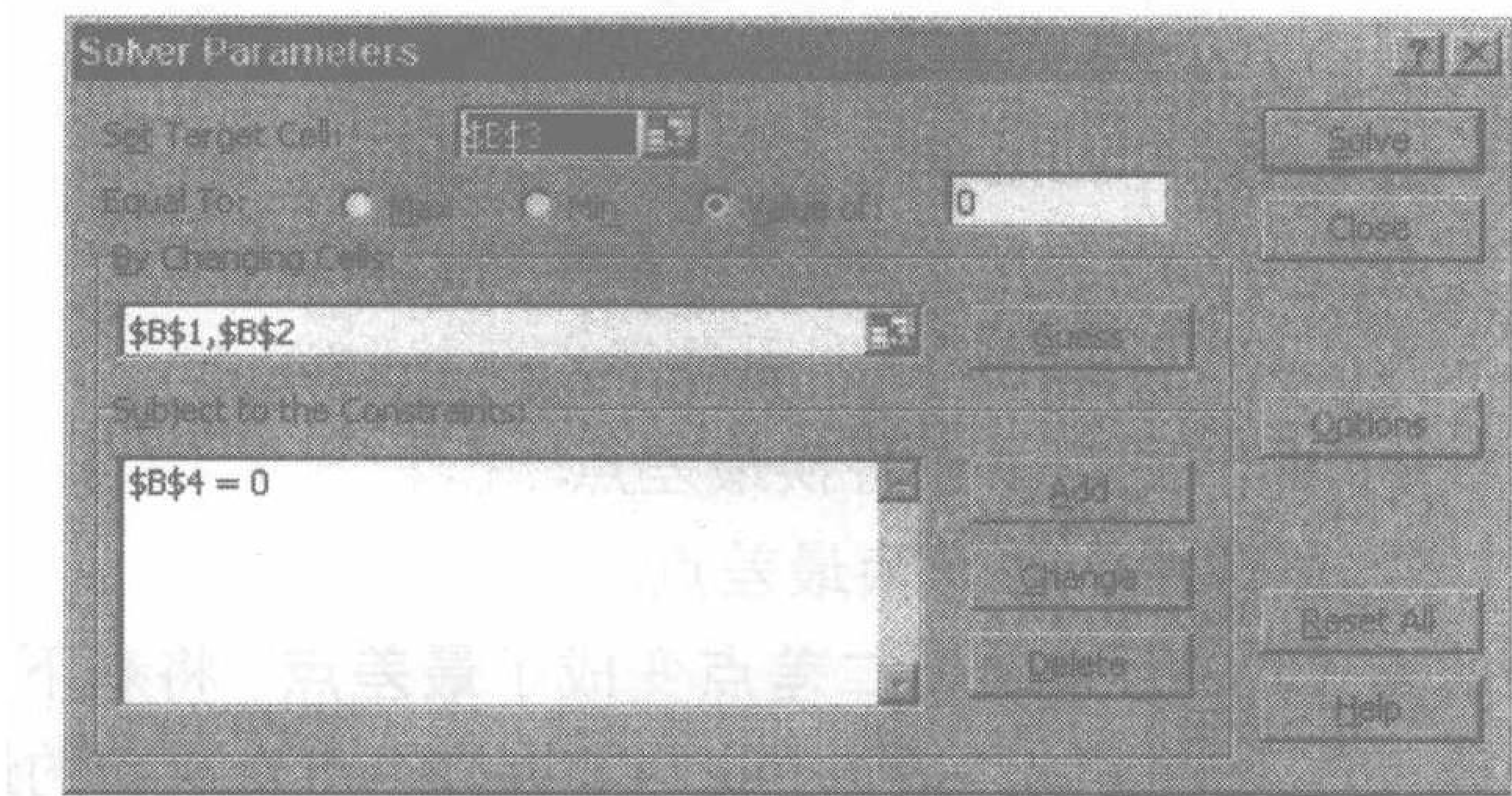
F.3 使用 Excel® 求解方程

作为最大化和最小化多变量函数的补充, 解算器还能解方程. 通过选择解算器对话框中的**值**选钮, 可以输入一个数值, 然后解算器通过手动改变**可变单元格**的值从而使**目标单元格**的值等于指定的值. 如果多于一个函数, 可以通过设置约束来实

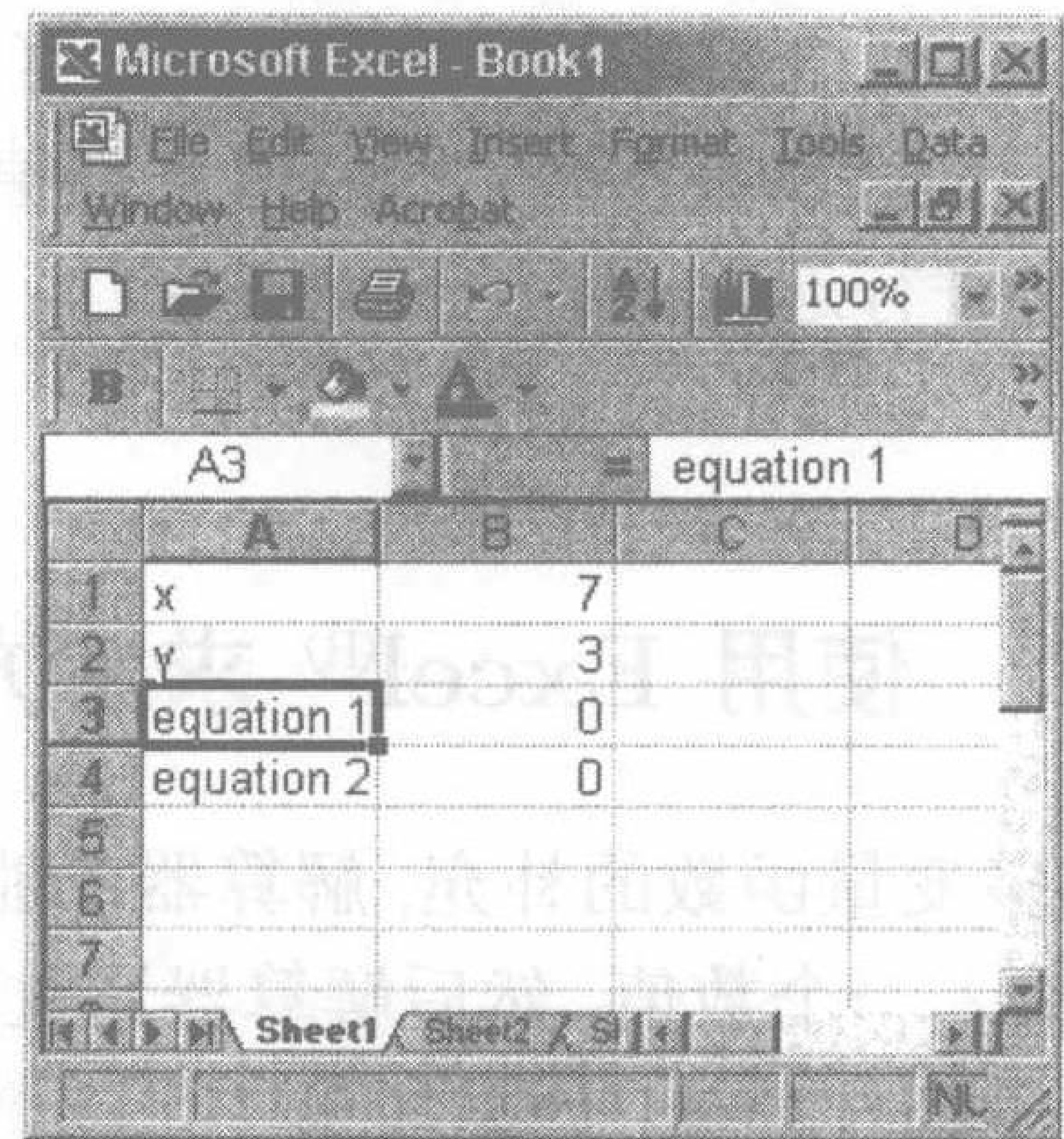
现. 下面的电子表格和解算器求解的对话框为求解以下两个方程: $x+y=10$ 和 $x-y=4$, 初值为 $x=8$ 和 $y=5$(为了说明初值可以不是任何一个方程的解).



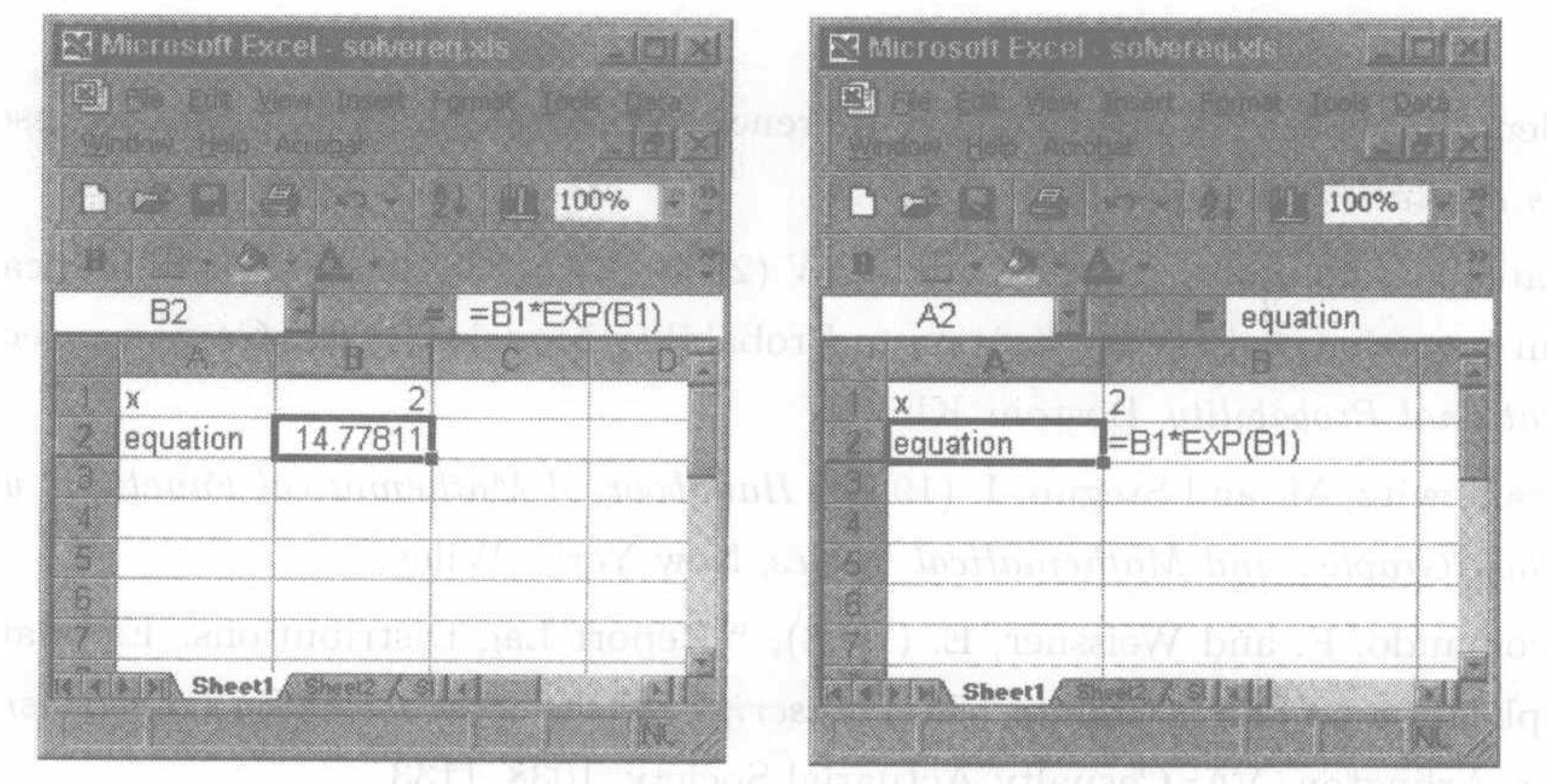
解算器求解的对话框为



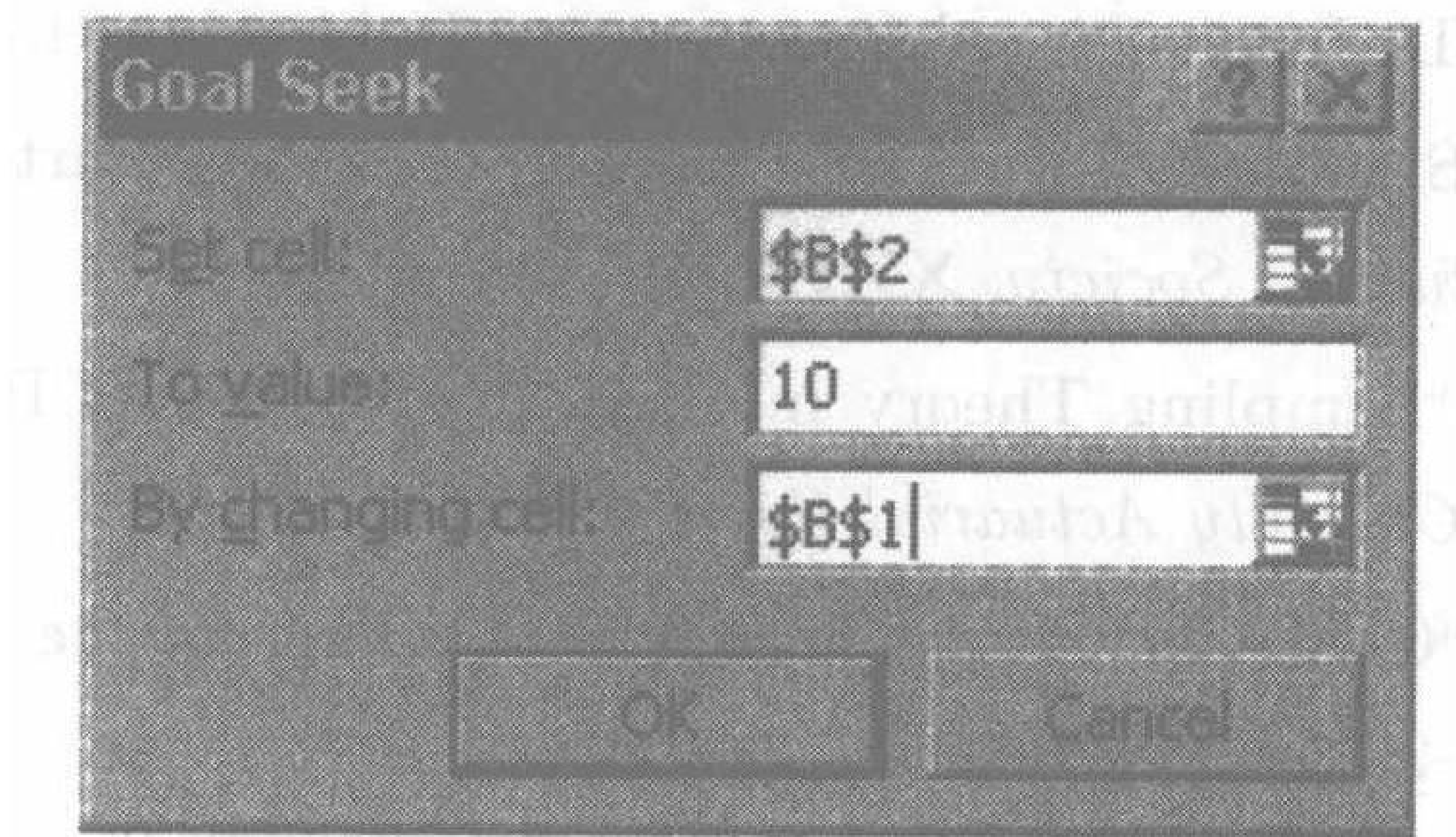
解为



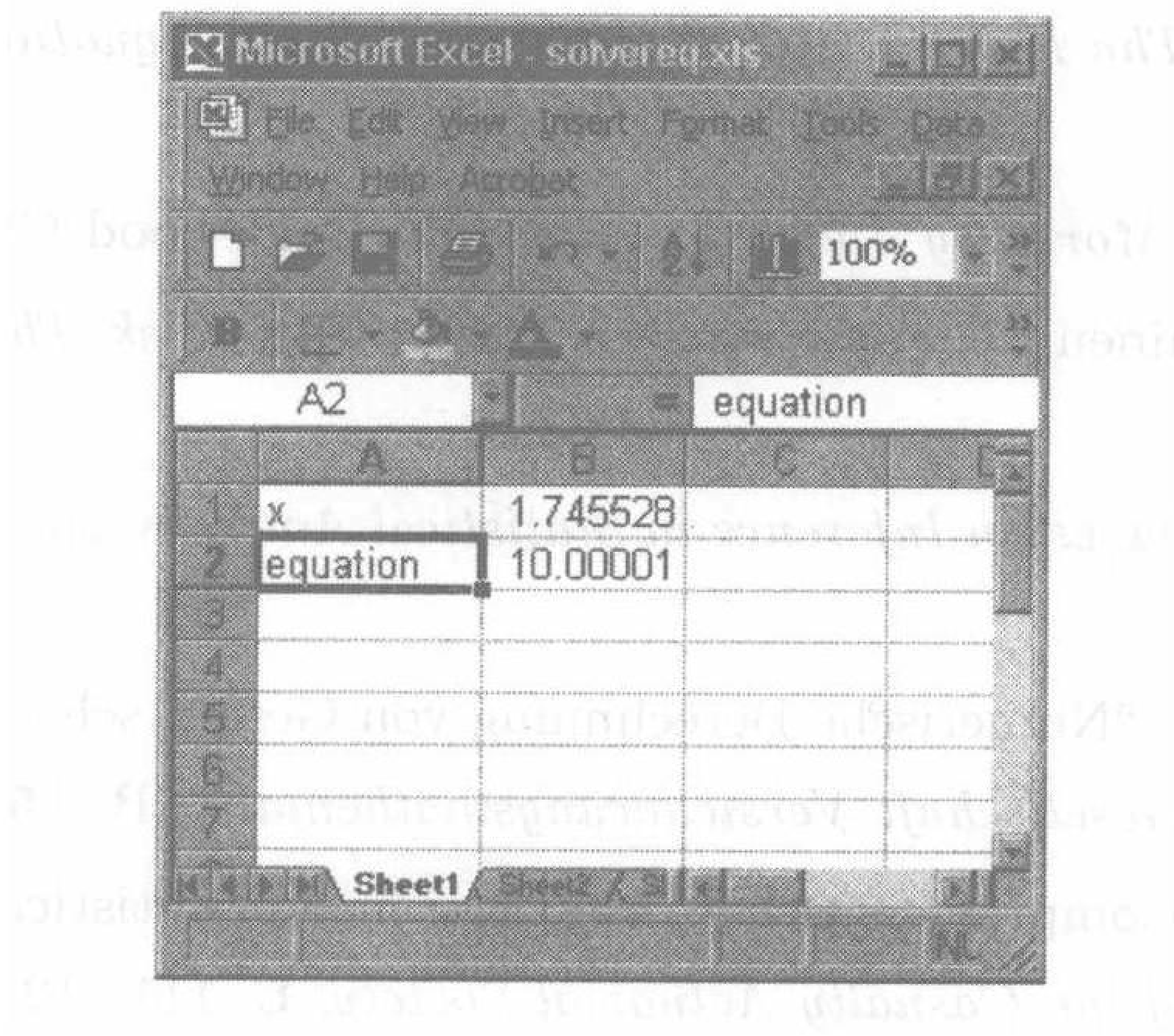
当只有一个方程和一个未知数时, Excel® 的单变量求解使用起来很简单. 大多数情况下在工具菜单中作为标准配备来安装. 假如希望求解 $xe^x = 10$, 下面简单的数据表格可用来求解初值 $x = 2$ 的情况.



单变量求解的对话框为



解为



参考文献

1. Aalen, O. (1978), "Nonparametric Inference for a Family of Counting Processes," *Annals of Statistics*, **6**, 701–726.
2. Abate, J., Choudhury, G., and Whitt, W.(2000), "An Introduction to Numerical Transform Inversion and Its Application to Probability Models," in W. Grassman, ed., *Computational Probability*, Boston: Kluwer.
3. Abramowitz, M. and Stegun, I. (1964), *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, New York: Wiley.
4. Accomando, F. and Weissner, E. (1988), " Report Lag Distributions: Estimation and Application to IBNR Counts," in *Transcripts of the 1988 Casualty Loss Reserve Seminar*, Arlington, VA: Casualty Actuarial Society, 1038–1133.
5. Arnold, B. (1983), *Pareto Distributions (Statistical Distributions in Scientific Work)*, Vol. 5, Fairland, MD: International Co-operative Publishing House.
6. Bailey, A. (1942), "Sampling Theory in Casualty Insurance, Parts I and II," *Proceedings of the Casualty Actuarial Society*, **XXIX**, 50–95.
7. Bailey, A. (1943), "Sampling Theory in Casualty Insurance, Parts III through VII," *Proceedings of the Casualty Actuarial Society*, **XXX**, 31–65.
8. Bailey, A. (1950), "Credibility Procedures," *Proceedings of the Casualty Actuarial Society*, **XXXVII**, 7–23 and 94–115.
9. Bailey, W. (1992), "A Method for Determining Confidence Intervals for Trend," *Transactions of the Society of Actuaries*, **XLIV**, 11–54.
10. Baker, C. (1977), *The Numerical Treatment of Integral Equations*, Oxford: Clarendon Press.
11. Batten, R. (1978), *Mortality Table Construction*, Englewood Cliffs, NJ: Prentice-Hall.
12. Beard, R., Pentikainen, T., and Pesonen, E. (1984), *Risk Theory*, 3rd ed., London: Chapman & Hall.
13. Berger, J. (1985), *Bayesian Inference in Statistical Analysis*, 2nd ed., New York: Springer-Verlag.
14. Bertram, J. (1981), "Numerische Berechnumg von Gesamtschadenvertei-lungen," *Blätter der deutschen Gesellschaft Versicherungsmathematik*, B. **15.2**, 175–194.
15. Bevan, J. (1963), "Comprehensive Medical Insurance—Statistical Analysis for Ratemaking," *Proceedings of the Casualty Actuarial Society*, **L**, 111–128.

16. Bowers, N., Gerber, H., Hickman, J., Jones, D., and Nesbitt, C. (1986), *Actuarial Mathematics*, Schaumburg IL: Society of Actuaries.

17. Brockett, P. (1991), "Information Theoretic Approach to Actuarial Science: A Unification and Extension of Relevant Theory and Applications," with discussion, *Transactions of the Society of Actuaries*, **XLIII**, 73–135.

18. Bühlmann, H. (1967), "Experience Rating and Credibility," *ASTIN Bulletin*, **4**, 199–207.

19. Bühlmann, H. (1970), *Mathematical Methods in Risk Theory*, New York: Springer-Verlag.

20. Bühlmann, H. and Straub, E. (1970), "Glaubwürdigkeit für Schadensätze (credibility for loss ratios)," *Mitteilungen der Vereinigung Schweizerischer Versicherungs-Mathematiker*, **70**, 111–133.

21. Carlin, B. and Klugman, S. (1993), "Hierarchical Bayesian Whitaker Graduation," *Scandinavian Actuarial Journal*, 183–196.

22. Carlin, B. and Louis, T. (2000), *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed., Boca Raton, FL: CRC Press.

23. Carriere, J. (1993), "Nonparametric Estimators of a Distribution Function Based on Mixtures of Gamma Distributions," *Actuarial Research Clearing House*, **1993.3**, 1–11.

24. Casualty Actuarial Society (1990), *Foundations of Casualty Actuarial Science*, Arlington, VA: Casualty Actuarial Society.

25. deAlba, E. (2002), "Bayesian Estimation of Outstanding Claim Reserves," *North American Actuarial Journal*, **6**, 1–20.

26. DePril, N. (1986), "On the Exact Computation of the Aggregate Claims Distribution in the Individual Life Model," *ASTIN Bulletin*, **16**, 109–112.

27. DePril, N. (1988), "Improved Approximations for the Aggregate Claims Distribution of a Life Insurance Portfolio," *Scandinavian Actuarial Journal*, 61–68.

28. DePril, N. (1989), "The Aggregate Claim Distribution in the Individual Model with Arbitrary Positive Claims," *ASTIN Bulletin*, **19**, 9–24.

29. Douglas, J. (1980), *Analysis with Standard Contagious Distributions*, Fairland, MD: International Co-operative Publishing House.

30. Dropkin, L. (1959), "Some Considerations on Automobile Rating Systems Utilizing Individual Driving Records," *Proceedings of the Casualty Actuarial Society*, **XLVI**, 165–176.

31. Efron, B. (1981), "Censored Data and the Bootstrap," *Journal of the American Statistical Association*, **76**, 321–319.

32. Efron, B. (1986), "Why Isn't Everyone a Bayesian?" *The American Statistician*, **40**, 1–11 (including comments and reply).

33. Efron, B. and Tibshirani, R. (1993), *An Introduction to the Bootstrap*, New York: Chapman & Hall.
34. Ericson, W. (1969), "A Note on the Posterior Mean of a Population Mean," *Journal of the Royal Statistical Society, Series B*, **31**, 332–334.
35. Feller, W. (1968), *An Introduction to Probability Theory and Its Applications*, Vol. 1, 3rd ed. rev., New York: Wiley.
36. Feller, W. (1971), *An Introduction to Probability Theory and Its Applications*, Vol. 2, 2nd ed., New York: Wiley.
37. Fisher, A. (1915), "Note on the Application of Recent Mathematical-Statistical Methods to Coal Mine Accidents, with Special Reference to Catastrophes in Coal Mines in the United States," *Proceedings of the Casualty Actuarial Society*, **II**, 70–78.
38. Fisz, M. (1963), *Probability Theory and Mathematical Statistics*, New York: Wiley.
39. Frees, E., Carriere, J., and Valdez, E. (1996), "Annuity Valuation with Dependent Mortality." *Journal of Risk and Insurance*, **63**, 229–261.
40. Frees, E. and Valdez, E. (1998) "Understanding Relationships Using Copulas," *North American Actuarial Journal*, **2**, 1–25.
41. Genest, C. (1987), "Frank's Family of Bivariate Distributions," *Biometrika*, **74**, 549–555.
42. Genest, C. and McKay, J. (1986), "The Joy of Copulas: Bivariate Distributions with Uniform Marginals," *The American Statistician*, **40**, 280–283.
43. Gerber, H. (1982), "On the Numerical Evaluation of the Distribution of Aggregate Claims and Its Stop-Loss Premiums," *Insurance: Mathematics and Economics*, **1**, 13–18.
44. Gerber, H. and D. Jones (1976), "Some Practical Considerations in Connection with the Calculation of Stop-Loss Premiums," *Transactions of the Society of Actuaries*, **XXVIII**, 215–231.
45. Gillam, W. (1992), "Parametrizing the Workers Compensation Experience Rating Plan," *Proceedings of the Casualty Actuarial Society*, **LXXIX**, 21–56.
46. Goovaerts, M. J. and Hoogstad, W. J. (1987), *Credibility Theory, Surveys of Actuarial Studies No. 4*, Rotterdam: Nationale-Nederlanden.
47. Guiahi, F. (2001), "Fitting to Loss Distributions with Emphasis on Rating Variables," *CAS Forum*, **Winter 2001**, 133–174.
48. Hachemeister, C. A. (1975), "Credibility for Regression Models with Application to Trend," in P. Kahn, ed., *Credibility: Theory and Applications*, New York: Academic Press, 129–163.
49. Harwayne, F. (1959), "Merit Rating in Private Passenger Automobile Liability Insurance and the California Driver Record Study," *Proceedings of the Casualty Actuarial Society*, **XLVI**, 189–195.

50. Hayne, R. (1994), "Extended Service Contracts," *Proceedings of the Casualty Actuarial Society*, **LXXXI**, 243–302.

51. Heckman, P. and G. Meyers (1983), "The Calculation of Aggregate Loss Distributions from Claim Severity and Claim Count Distributions," *Proceedings of the Casualty Actuarial Society*, **LXX**, 22–61.

52. Herzog, T. (1999), *Introduction to Credibility Theory*, 3rd ed., Winsted, CT: ACTEX.

53. Herzog, T. and Lord, G. (2002), *Applications of Monte Carlo Methods to Finance and Insurance*, Winsted, CT: ACTEX.

54. Herzog, T. and Laverty, J. (1995), "Experience of Refinanced FHA Section 203(b) Single Family Mortages," *Actuarial Research Clearing House*, **1995.1**, 97–129.

55. Hewitt, C., Jr. (1967), "Loss Ratio Distributions—A Model," *Proceedings of the Casualty Actuarial Society*, **LIV**, 70–88.

56. Hewitt, C., Jr., and Lefkowitz, B. (1979), "Methods for Fitting Distributions to Insurance Loss Data," *Proceedings of the Casualty Actuarial Society*, **LXVI**, 139–160.

57. Hipp, G. (1938), " Special Funds Under the New York Workmen's Compensation Law," *Proceedings of the Casualty Actuarial Society*, **XXIV**, 247–275.

58. Hogg, R. and Craig, A. (1978), *Introduction to Mathematical Statistics*, 4th ed., New York: Macmillan.

59. Hogg, R. and Klugman, S. (1984), *Loss Distributions*, New York: Wiley.

60. Holgate, p. (1970), "The Modality of Some Compound Poisson Distributions," *Biometrika*, **57**, 666–667.

61. Holler, K., Sommer, D., and Trahair, G. (1999), "Something Old, Some thing New in Classification Ratemaking with a Novel Use of Generalized Linear Models for Credit Insurance," *CAS Forum*, **Winter 1999**, 31–84.

62. Hossack, I., Pollard, J., and Zehnwirth, B. (1983), *Introductory Statistics with Applications in General Insurance*, Cambridge: Cambridge University Press.

63. Hougaard, P. (2000), *Analysis of Multivariate Survival Data*, New York: Springer-Verlag.

64. Hutchinson, T. and Lai, C. (1990), *Continuous Bivariate Distributions, Emphasizing Applications*, Adelaide: Rumsby.

65. Hyndman, R. and Fan, Y. (1996), "Sample Quantiles in Statistical Packages," *The American Statistician*, **50**, 361–365.

66. Jewell, W. (1974), "Credibility Is Exact Bayesian for Exponential Families," *ASTIN Bulletin*, **8**, 77–90.

67. Johnson, N., Kotz, S., and Balakrishnan, N. (1994), *Continuous Univariate Distributions*, Vol. 1, 2nd ed., New York: Wiley.

68. Johnson, N., Kotz, S., and Balakrishnan, N. (1995), *Continuous Univariate Distributions*, Vol. 2, 2nd ed., New York: Wiley.
69. Johnson, N., Kotz, S., and Kemp, A. (1993), *Univariate Discrete Distributions*, 2nd ed., New York: Wiley.
70. Kaplan, E. and Meier, P. (1985), "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association*, **53**, 457–481.
71. Karlin, S. and Taylor, H. (1975), *A First Course in Stochastic Processes*, 2nd ed., New York: Academic Press.
72. Karlin, S. and Taylor, H. (1981), *A Second Course in Stochastic Processes*, New York: Academic Press.
73. Kleiber, C. and Kotz, S. (2003), *Statistical Size Distributions in Economics and Actuarial Sciences*, New York: Wiley.
74. Klein, J. and Moeschberger, M. (1997), *Survival Analysis, Techniques for Censored and Truncated Data*, New York: Springer-Verlag.
75. Klugman, S. (1981), "On the Variance and Mean Squared Error of Decrement Estimators," *Transactions of the Society of Actuaries*, **XXXIII**, 301–311.
76. Klugman, S. (1987), "Credibility for Classification Ratemaking Via the Hierarchical Linear Model," *Proceedings of the Casualty Actuarial Society*, **LXXIV**, 272–321.
77. Klugman, S. (1992), *Bayesian Statistics in Actuarial Science with Emphasis on Credibility*, Boston: Kluwer.
78. Klugman, S. and Parse, A. (1999), "Fitting Bivariate Distributions with Copulas," *Insurance: Mathematics and Economies*, **24**, 139–148.
79. Kornya, P. (1983), "Distribution of Aggregate Claims in the Individual Risk Model," *Transactions of the society of Actuaries*, **XXXV**, 837–858.
80. Kotz, S., Balakrishnan, N., and Johnson, N. (2000), *Continuous Multivariate Distributions*, Vol. 1, Models and Applications, New York: Wiley.
81. Lawless, J. (2003), *Statistical Models and Methods for Lifetime Data*, 2nd ed., New York: Wiley.
82. Lemaire, J. (1995), *Automobile Insurance: Actuarial Models*, 2nd ed., Boston: Kluwer.
83. Lindley, D. (1987), "The Probability Approach to the Treatment of Uncertainty in Artificial Intelligence and Expert Systems," *Statistical Science*, **2**, 17–24 (also related articles in that issue).
84. London, D. (1985), *Graduation: The Revision of Estimates*, Winsted, CT: ACTEX.
85. London, D. (1988), *Survival Models and Their Estimation*, 3rd ed., Winsted, CT: ACTEX.
86. Longley-Cook, L. (1958), "The Employment of Property and Casualty Actuaries," *Proceedings of the Casualty Actuarial Society*, **XLV**, 9–10.

87. Longley-Cook, L. (1962), "An Introduction to Credibility Theory," *Proceeding of the Casualty Actuarial Society*, **XLIX**, 194–221.
88. Luong, A. and Doray, L. (1996), "Goodness of Fit Test Statistics of the Zeta Family," *Insurance: Mathematics and Economics*, **10**, 45–53.
89. Mardia, K. (1970), *Families of Bivariate Distributions*, London: Griffin.
90. McCullagh, P. and Nelder, J. (1989), *Generalized Linear Models*, New York: Chapman & Hall.
91. Meyers, G. (1984), "Empirical Bayesian Credibility for Workers'Compensation Classification Ratemaking," *Proceedings of the Casualty Actuarial Society*, **LXXI**, 96–121.
92. Meyers, G. (1994), "Quantifying the Uncertainty in Claim Severity Estimates for an Excess Layer When Using the Single Parameter Pareto," *Proceeding of the Casualty Actuarial Society*, **LXXXI**, 91–122(including discussion).
93. Mildenhall, S. (1999), "A Systematic Relationship between Minimum Bias and Generalized Linear Models," *Proceedings of the Casualty Actuarial Society*, **LXXXVI**, 393–487.
94. Miller, M.(1949), *Elements of Graduation*, Philadelphia: The Actuarial Society of America and the American Institute of Actuaries.
95. Moore, D. (1986), "Tests of Chi-Squared Type," in D'Agostino, R. and Stephens, M., eds., *Goodness-of-Fit Techniques*, New York: Marcel Dekker, 63–95.
96. Mowbray, A. H. (1914), "How Extensive a Payroll Exposure Is Necessary to Give a Dependable Pure Premium?" *Proceedings of the Casualty Actuarial Society*, **I**, 24–30.
97. Murphy, K., Brockman, M., and Lee, P. (2000), "Using Generalized Linear Models to Build Dynamic Pricing Systems for Personal Lines Insurance," *CAS Forum*, **Winter 2000**, 107–140.
98. Nelder, J. and Mead, U. (1965), "A Simples Method for Function Minimization," *The Computer Journal*, **7**, 308–313.
99. Nelson, W. (1972), "Theory and Applications of Hazard Plotting for Censored Failure Data," *Technometrics*, **14**, 945–965.
100. Norberg, R. (1979), "The Credibility Approach to Experience Rating," *Scandinavian Actuarial Journal*, 181–221.
101. Ntzoufras, I. and Dellaportas, P. (2002), "Bayesian Modeling of Oustanding Liabilities Incorporating Claim Count Uncertainty," *North American Actuarial Journal*, **6**, 113–128.
102. Patrik, G. (1980), "Estimating Casualty Insurance Loss Amount Distributions," *Proceedings of the Casualty Actuarial Society*, **LXVII**, 57–109.
103. Panjer, H. and Lutek, B. (1983), "Practical Aspects of Stop-Loss Calculations," *Insurance: Mathematics and Economics*, **2**, 159–177.

104. Panjer, H. and Wang, S. (1993), "On the Stability of Recursive Formulas," *ASTIN Bulletin*, **23**, 227–258.

105. Panjer, H. and Willmot, G. (1986), "Computational Aspects of Recursive Evaluation of Compound Distributions," *Insurance: Mathematics and Economics*, **5**, 113–116.

106. Panjer, H. and Willmot, G. (1992), *Insurance Risk Models*, Chicago: Society of Actuaries.

107. Press, W., Flannery, B., Teukolsky, S., and Vetterling, W. (1988), *Numerical Recipes in C*, Cambridge: Cambridge University Press.

108. Rao, C. (1965), *Linear Statistical Inference and Its Applications*, New York: Wiley.

109. Rioux, J. and Klugman S. (2003), "Toward a Unified Approach to Fitting Loss Models," working paper.

110. Ripley, B. (1987), *Stochastic Simulation*, New York: Wiley.

111. Robertson, J. (1992), "The Computation of Aggregate Loss Distributions," *Proceedings of the Casualty Actuarial Society*, **LXXIX**, 57–133.

112. Rohatgi, V. (1976), *An Introduction to Probability Theory and Mathematical Statistics*, New York: Wiley.

113. Rolski, T., Schmidli, H., Schmidt, V., and Teugels, J. (1999), *Stochastic Processes for Insurance and Finance*, Chichester: Wiley.

114. Ross, S. (1996), *Stochastic Processes*, 2nd ed., New York: Wiley.

115. Ross, S. (2002), *Simulation*, 3rd ed., San Diego: Academic Press.

116. Ross, S. (2003), *Introduction to Probability Models*, 8th ed., San Diego: Academic Press.

117. Schoenberg, I. (1964), "Spline Functions and the Problem of Graduation," *Proceedings of the National Academy of Science*, **52**, 947-950.

118. Scollnik, D. (2001), "Actuarial Modeling with MCMC and BUGS," *North American Actuarial Journal*, **5**, 96–124.

119. Scollnik, D. (2002), "Modeling Size-of-Loss Distributions for Exact Data in WinBUGS," *Journal of Actuarial Practice*, **10**, 193–218.

120. Self, S. and Liang, K. (1978), "Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions," *Journal of the American Statistical Association*, **82**, 605–610.

121. Schwarz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, **6**, 461–464.

122. Simon, L. (1961), "Fitting Negative Binomial Distributions by the Method of Maximum Likelihood," *Proceedings of the Casualty Actuarial Society*, **XLVIII**, 45–53.

123. Society of Actuaries Committee on Actuarial Principles (1992), "Principles of Actuarial Science," *Transactions of the Society of Actuaries*, **XLIV**, 565–628.

124. Society of Actuaries Committee on Actuarial Principles (1995), "Principles Regarding Provisions for Life Risks," *Transactions of the Society of Actuaries*, **XLVII**, 775–793.

125. Stephens, M. (1986), "Tests Based on EDF Statistics," in D'Agostino, R. and Stephens, M., eds., *Goodness-of-Fit Techniques*, New York: Marcel Dekker, 97–193.

126. Sundt, B. (1986), Special issue on credibility theory, *Insurance: Abstracts and Reviews*, **2**.

127. Sundt, B. (1999), *An Introduction to Non-Life Insurance Mathematics*, 4th ed., Mannheim: University of Mannheim Press.

128. Thyrion, P. (1961), "Contribution a l'Etude du Bonus pour non Sinstre en Assurance Automobile," *ASTIN Bulletin*, **1**, 142–162.

129. Tijms, H. (1994), *Stochastic Models—An Algorithmic Approach*, Chichester: Wiley.

130. Tröbliger, A. (1961), "Mathematische Untersuchungen zur Beitragsruchgewahr in der Kraftfahrversicherung," *Blatter der Deutsche Gesellschaft fur Versicherungsmathematik*, **5**, 327–348.

131. Tukey, J. (1962), "The future of data analysis," *Annals of Mathematical Statistics*, **33**, 1–67.

132. Venter, G. (1983), "Transformed Beta and Gamma Distributions and Aggregate Losses," *Proceedings of the Casualty Actuarial Society*, **LXX**, 156–193.

133. Verrall, R. (1990), "Bayes and Emipirical Bayes Estimation for the Chain Ladder Method," *ASTIN Bulletin*, **20**, 217–243.

134. Walters. F., Parker, L., Morgan, S., and Deming, S. (1991), *Sequential Simplex Optimization*, Boca Raton, FL: CRC Press.

135. Waters, H. R. (1993), *Credibility Theory*, Edinburgh: Department of Actuarial Mathematics & Statistics, Heriot-Watt University.

136. Whitney, A. W. (1918), "The Theory of Experience Rating," *Proceedings of the Casualty Actuarial Society*, **IV**, 274–292.

137. Willmot, G. (1998), "On a Class of Approximations for Ruin and Waiting Time Probabilities," *Operations Research Letters*, **22**, 27–32.

索　　引